

Depression Detection Using Hearing-Aid Integrated Sensors

Master's Thesis in Artificial Intelligence

submitted
by

Maximilian Renz

born 11.06.1999 in Dillingen a.d. Donau

Written at

Machine Learning and Data Analytics Lab
Department Artificial Intelligence in Biomedical Engineering
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

Advisors: Ann-Kristin Seifer, M. Sc., Dr.-Ing. Robert Richer, Lydia Rupp, M. Sc., Prof. Dr. Matthias Berking, Prof. Dr. Björn Eskofier

Started: 09.12.2024

Finished: 10.06.2025

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Die Richtlinien des Lehrstuhls für Bachelor- und Masterarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Erlangen, den 6. Juni 2025

Übersicht

Depression gehört weltweit zu den häufigsten psychischen Erkrankungen und beeinträchtigt zahlreiche Lebensbereiche der Betroffenen. Häufig bleibt sie unerkannt – unter anderem aufgrund von Stigmatisierung, eingeschränktem Zugang zu klinischen Angeboten und der Abhängigkeit von selbstberichteten Symptomen – was ihre negativen Auswirkungen zusätzlich verstärkt. Derzeit erfolgt die Diagnose durch geschulte Fachkräfte mittels strukturierter Interviews und standardisierter Fragebögen. Mit den jüngsten Fortschritten im Bereich der digitalen Gesundheit gewinnen objektive digitale Biomarker zunehmend an Bedeutung, da sie eine passive, unauffällige und kosteneffiziente Erfassung verhaltensbezogener Merkmale zur Depressionsdiagnostik ermöglichen.

Diese Arbeit stellt eine neuartige Methode zur Erkennung von Depressionen vor, die auf Sensoren basiert, die bereits in Hörgeräten integriert sind. Diese Lösung ist sowohl diskret als auch alltagstauglich und ermöglicht ein kontinuierliches Monitoring, da das zugrundeliegende Gerät bereits fester Bestandteil des täglichen Lebens vieler Menschen ist. Hörgeräte verfügen typischerweise über ein Mikrofon und eine inertielle Messeinheit, wodurch sich multi-modale Daten – insbesondere Sprache, Gang und Kopfbewegungen – ohne zusätzliches Equipment erfassen lassen.

Zur Validierung dieses Ansatzes wurde eine klinische Studie mit 25 Teilnehmenden durchgeführt, von denen letztlich 21 (zwölf gesunde und neun depressive Personen gemäß einem PHQ-8-Wert ≥ 10) in die Analyse einbezogen wurden. Sprach- und Bewegungsdaten wurden ausschließlich über in Hörgeräten integrierte Sensoren erfasst – im Rahmen halbstrukturierter Interviews sowie bei Geh- und Bewegungstests, die auf natürliche Weise emotionale und verhaltensbezogene Reaktionen hervorrufen sollten. Die akustische Analyse basierte auf einem reduzierten, klinisch relevanten Set paralinguistischer Sprachmerkmale sowie emotionsbezogener Features. Die Bewegungsanalyse umfasste spatio-temporale Gangmerkmale, generische Zeitreihenkennwerte und expertengestützte Merkmale.

Sowohl statistische Verfahren als auch klassische machine learning-Modelle, darunter logistische Regression, *k-nearest neighbors*, *random forest*, *support vector machine* und *extreme gradient boosting*, wurden für verschiedene Merkmalskombinationen eingesetzt. Die Ergebnisse zeigen, dass charakteristische Verhaltensmarker, wie sie in der Literatur beschrieben werden, grundsätzlich auch mit am Ohr getragenen Sensoren reproduziert werden können. Sowohl Sprach- als auch Gangdaten ermöglichten unabhängig voneinander eine zuverlässige Unterscheidung zwischen gesunden und depressiven Personen. Besonders hervorzuheben ist, dass die Kombination der Modalitäten die Klassifikationsgenauigkeit weiter verbesserte – mit einer finalen Genauigkeit von 85 % mit *random forest* bei Nutzung aller verfügbaren Datenquellen.

Diese Arbeit belegt die Machbarkeit einer kontinuierlichen, datenschutzfreundlichen Erkennung von Depressionen mithilfe von Geräten, die bereits in den Alltag integriert sind. Die Ergebnisse unterstreichen das Potenzial von *Earables* – insbesondere Hörgeräten – als unauffällige, multi-modale Plattformen für die mentale Gesundheitsüberwachung und schaffen die Grundlage für zukünftige groß angelegte Studien sowie praxisnahe Anwendungen im Bereich der mentalen Gesundheit.

Abstract

Depression is one of the most prevalent mental illnesses worldwide, negatively impacting numerous aspects of individuals' lives. It often remains undiagnosed due to stigma, limited access to clinical services, and the reliance on self-reported symptoms, worsening its effects. Currently, depression is diagnosed by trained professionals using structured interviews and questionnaires. With recent advances in digital health, objective digital biomarkers are emerging as promising tools for detecting depression in a passive, unobtrusive, and cost-effective manner by monitoring behavioral markers.

This thesis proposes a novel method for depression detection using sensors inherently integrated into hearing aids (HAs). This solution is both discreet and practical, offering a pathway toward continuous monitoring by leveraging a device already embedded in users' daily lives. HAs naturally incorporate both a microphone and an inertial measurement unit (IMU), enabling the collection of multi-modal data – including voice, gait, and head motion – without additional hardware.

To support this approach, a clinical study was conducted with 25 participants, ultimately including 21 (twelve healthy, nine depressed as defined by PHQ-8 scores ≥ 10). Voice and motion data were captured exclusively via HA-embedded sensors during semi-structured interviews and walking tasks designed to elicit behavioral and emotional responses in naturalistic settings. Audio features included a minimalistic set of clinically relevant paralinguistic markers and emotion-related features. Motion analysis incorporated spatio-temporal gait features, generic time-series descriptors, and expert-designed markers. Statistical analyses and a range of classical machine learning (ML) models, including logistic regression (LR), k-nearest neighbors (kNN), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGBoost), were applied across various combinations of feature sets.

The results demonstrate that characteristic behavioral markers of depression reported in the literature can generally be reproduced using ear-worn sensors. Both speech and gait modalities independently enabled meaningful classification between healthy and depressed participants. Notably, the fusion of features across modalities improved performance further, achieving a classification accuracy of 85 % with RF when combining all available data domains.

This work highlights the feasibility of continuous, privacy-preserving depression detection using devices already integrated into everyday life. The findings support the potential of earables, especially hearing aids, as viable platforms for unobtrusive, multi-modal mental health assessment and establish the foundation for future large-scale studies and real-world applications in mental health monitoring.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Purpose of Work	3
1.3	Outline	4
2	Background	5
2.1	Depression	5
2.2	Speech	6
2.2.1	Generation	6
2.2.2	Pathology	8
2.3	Gait and Motion	9
2.3.1	Walking Patterns	9
2.3.2	Motion Analysis	10
2.3.3	Gait Analysis	11
2.4	Machine Learning	12
2.4.1	Logistic Regression (LR)	13
2.4.2	k-Nearest Neighbors (kNN)	13
2.4.3	Random Forest (RF)	14
2.4.4	Support Vector Machine (SVM)	15
2.4.5	Extreme Gradient Boosting (XGBoost)	16
2.4.6	Transformer	17
3	Related Work	19
3.1	Voice-Based Depression Detection	19
3.2	Gait and Motion Analysis in Depression Detection	22
3.3	Fusion	25

3.4	Ear-Worn Sensors	25
4	Data Collection	27
4.1	Study Design	27
4.2	Procedure	29
5	Methods	31
5.1	Voice Analysis	31
5.1.1	Pre-Processing	31
5.1.2	Feature Extraction	32
5.2	Gait and Motion Analysis	33
5.2.1	Pre-Processing	33
5.2.2	Feature Extraction	34
5.3	Evaluation	35
5.3.1	Statistical Analysis	35
5.3.2	Classification Experiments	37
6	Results	39
6.1	Statistical Analysis	39
6.2	Classification Experiments	45
7	Discussion	51
8	Conclusion and Outlook	55
	List of Figures	57
	List of Tables	59
	Bibliography	61
A	Study Protocol	81
B	Study Procedure	97
C	Hyperparameter Search Space	107
D	Acronyms	109

Chapter 1

Introduction

1.1 Motivation

Depressive disorders are among the most prevalent mental illnesses. According to estimates from the World Health Organization, depression affects more than 280 million individuals globally, including 5% of the adult population [WHO23]. The presence of symptoms such as sadness, subdued mood, and listlessness is pervasive and forms part of normal life, to which significant functions in coping with stressful events can be attributed. However, depression in the clinical sense can arise if the impairments exceed normal human unhappiness in terms of intensity, duration, situational inappropriateness, and extent of suffering. In addition, a wide range of other symptoms may be present, which vary significantly in both kind and intensity. Affected individuals often report feelings of desperation or a sense of inner emptiness, as well as a negative attitude towards themselves and the future. These symptoms may manifest in various forms, including social withdrawal, insomnia, and a loss of appetite [Ber12]. Depression to such an extent represents a considerable burden for those affected, entailing a loss of quality of life and a reduction in life expectancy [Jia15].

The failure to seek treatment in a timely manner is a phenomenon that can be ascribed to various factors such as stigma, lack of awareness, or inadequate provision of mental health services. It has the potential to result in the development of more severe, chronic, and treatment-resistant conditions. The absence of treatment for mental disorders can further contribute to social and occupational challenges, unstable interpersonal relationships, and an elevated risk of suicide [Wan07]. Conversely, timely intervention can significantly improve outcomes [Het08], with earlier treatment being associated with higher remission rates [Buk13], thereby enhancing mental and social functioning [Cou97].

In order to establish a valid diagnosis of depression and differentiate it from normal mood fluctuations and other affective disorders, it is imperative to inquire about the full spectrum of depressive symptoms [Ber12]. This is typically undertaken by a trained professional (e.g. physicians, psychiatrists, psychologists) using structured interviews and standardized questionnaires such as the Hamilton Rating Scale for Depression (HAM-D) [Ham80], the Beck Depression Inventory-II (BDI-II) [Bec96], and the Patient Health Questionnaire-8 (PHQ-8) [Kro09]. However, these assessments can only be performed when patients seek medical care from these professionals.

A field of research is currently emerging that explores alternative approaches for the detection of depression through the assessment of objective digital biomarkers. These biomarkers are considered digital due to their collection via pervasive sensors, offering scalability, unobtrusive monitoring, temporal sensitivity, and cost-effectiveness [Jac19]. They encompass parameters such as voice, activity, gait, sleep time, and numerous additional parameters. Notably, these biomarkers show significant potential as indicators of high-level behavioral characteristics associated with depression and anxiety [Moh17]. Voice, for example, is a frequently studied biomarker [Wu23a; Low20], given that individuals with depression exhibit notable alterations in their speech patterns. These include, among others, a reduction in intensity, a decrease in fundamental frequency deviation or pitch range, and a slowing of speech rate [Hol80]. These non-verbal vocal characteristics, known as paralinguistic features, enabled the development of numerous automated speech analysis techniques using machine learning [Wu23a]. Another objective digital biomarker that can be indicative of depression is gait, including reduced gait speed and longer double limb support. Additionally, aspects such as slumped posture and impaired dynamic balance have been shown to reflect the impact of depression on overall body movement and stability [Fel20]. Furthermore, a variety of other motion features have already been shown to be useful for assessing the influence of stress on body posture and movement [Ric24], which may also serve as promising indicators for detecting depression.

While voice, gait and other motion features are suitable for depression detection, collecting these data present significant challenges, because sensors must obtain highly personal information over an extended period on a daily basis. Conventional wearable systems designed for voice analysis and movement tracking often require dedicated devices that can be quite obtrusive, limiting their integration into daily life. In this context, earables, such as hearing aids (HAs) and advanced headphones, offer a compelling alternative due to their seamless integration into everyday routines. Earables are not only unobtrusive but also socially acceptable, discreet, privacy-preserving, and non-stigmatizing [Röd22]. This is especially important, when dealing with depression, which is already associated with stigmatization [Wan07].

HAs, in particular, are evolving into multifunctional health devices. In addition to their primary function, they are increasingly equipped with built-in Bluetooth audio and inertial measurement unit (IMU) sensors, enabling new capabilities. These include direct streaming of phone calls or music from smartphones, as well as the classification of acoustic environments, which allows for automatic adjustments to background noise based ambient sounds and subtle head movements. This technological advancement also facilitates the continuous and unobtrusive collection of voice, gait, and other movement data, making them especially well-suited for monitoring mental health and physical well-being, including the automatic detection of falls and social engagement patterns [Bho21]. This positions earables as a promising tool for monitoring behavioral patterns associated with depression. By leveraging these naturally integrated features, earables offer a compelling solution to address the practical limitations of conventional wearable systems in depression detection.

1.2 Purpose of Work

Although a wide range of depression detection approaches based on speech, gait, and motion have already been evaluated and demonstrated promising results, they frequently rely on a combination of audio-visual cues [Val16; Nas16] or require the use of external sensing devices [Zha19; Li21]. Studies employing passive sensing techniques often require the participants to wear one or more additional sensors [OBr17; Jun22] or incorporate private personal data from smartphones and wristbands [Nar20].

To address these limitations, this work proposes a novel method for depression detection using sensors already integrated into hearing aids, offering a discreet and practical solution. HAs naturally integrate both a microphone and an IMU, enabling efficient and continuous acquisition of multi-modal data. By utilizing a device that is already incorporated into the daily life for many individuals, this approach offers a non-intrusive means of monitoring mental health through multiple data modalities.

To reach this goal, a study will be conducted involving approximately 20 participants, evenly split between individuals with depression and healthy controls. Data collection will be carried out exclusively through the utilization of HA-integrated sensors, specifically the inertial sensor and the microphone. The PHQ-8 will be used to assess the presence and severity of depression. Voice data will be gathered through a semi-structured interview, while gait-related data will be captured by the IMU using a 6-minute walking test. Additionally, head movements will be recorded in different situations to provide further insight into the participants' activity and movement.

To predict the depression score, different algorithmic approaches will be implemented and evaluated. Voice features will be extracted using both openSMILE [Eyb10] and wav2vec 2.0, a deep learning model for emotion recognition [Wag23]. Gait features will be derived using EarGait [Sei23], while motion characteristics will be analyzed using both generic time-series features from tsfresh [Chr18] and expert-designed features as proposed in [Ric24]. To assess the relative importance and predictive power of each data modality, various feature combinations will be explored. For classification, several machine learning models will be applied, including logistic regression (LR), k-nearest neighbors (kNN), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGBoost). These models will be evaluated to determine their effectiveness in distinguishing between healthy and depressed individuals.

1.3 Outline

In Chapter 2, this thesis first provides a comprehensive overview of fundamentals concerning human speech and gait, along with the machine and deep learning techniques employed within the scope of this thesis. Next, Chapter 3 provides an overview of the existing research in this field. The details of the underlying study are presented in Chapter 4. Chapter 5 will elaborate the methods, which can be divided into data pre-processing, feature extraction, and the machine learning and deep learning pipeline of both modalities as well as their fusion. The results will be laid out in Chapter 6 and subsequently, the findings will be discussed in Chapter 7. The thesis concludes with Chapter 8, in which the key findings and limitations of this work are summarized and an outlook about future work is provided.

Chapter 2

Background

2.1 Depression

Depression is an affective disorder characterized by the presence of depressive symptoms in the absence of manic symptoms [Ber12]. It is typically categorized into different main forms, such as a single depressive episode, recurrent depressive disorder – defined by repeated depressive episodes – and persistent mood disorder, which may last for many years or even the majority of an individual's adult life [WHO04].

In addition to core symptoms such as persistently depressed mood, anhedonia, loss of interest, and listlessness, a variety of secondary criteria commonly accompany depression. These include affective symptoms, such as sadness, despair, or feeling of inner emptiness, and cognitive symptoms, for instance negative thoughts about yourself and problems with concentration. Furthermore, there are motivational-behavioral characteristics, such as social withdrawal or slowed speech and motor skills, and somatic symptoms, for instance changes in appetite, loss of energy, and insomnia [Ber12].

Depression significantly impairs functioning across nearly all domains of daily life, including work, home management, activities, and social interactions. These symptoms effect each of these domains differently [Fri14]. In the workplace, affected individuals often experience reduced productivity, increased absenteeism, and heightened vulnerability to unemployment [Ler08]. Social functioning is also commonly affected, with individuals reporting diminished empathy, reduced interest in social engagement, and difficulty in forming or maintaining relationships [Kup16]. This is particularly concerning given that a strong social network has been shown to have a protective effect against depression [San15].

Given the broad spectrum of symptoms, it is essential to assess the full range during diagnosis.

To this end, clinicians employ structured diagnostic interviews [Pfo97] and standardized assessment tools such as the HAM-D [Ham80], the BDI-II [Bec96], and the PHQ-8 [Kro09] are used.

Once diagnosed, depression can often be treated effectively through psychotherapy. Evidence-based approaches are also available for the treatment of chronic depression, relapse prevention, and specific patient subgroups. For severe cases, combination treatments involving both psychotherapy and pharmacotherapy are recommended and have demonstrated particular efficacy [Ber12].

2.2 Speech

2.2.1 Generation

Humans use their voice to communicate and to convey meaning through linguistic (words and grammar) and paralinguistic (pitch and tone) information. It is generated by our vocal tract, which comprises the lips, teeth, mouth, tongue, and larynx [Ram21].

The process of speak generation generally commences with airflow generated by the lungs. During quiet breathing, the respiratory muscles ensure regularly alternating inhalation and exhalation. However, during speech, this pattern changes towards a prolonged expiratory phase, with a consistent lung pressure being maintained [Hon08]. As seen in Figure 2.1, the air travels further through the larynx, more specifically through the glottis, which is the space formed by the vocal

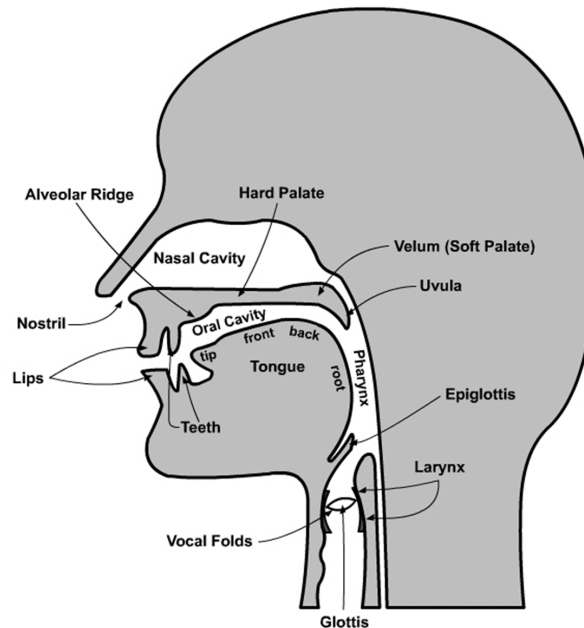


Figure 2.1: Overview of the vocal tract, adapted from from [Do 25].

folds. Surrounded by multiple muscles and cartilages, these vocal folds can be modulated in both position and properties through relative movement to each other [Zha16].

The initiation of the phonation process is characterized by the adduction of the vocal folds, which closes the glottis and generates subglottal pressure. If a certain threshold is exceeded, the glottis opens again, which initiates a wave-like, self-sustained motion of the vocal folds into a cycle of opening and closing [Zha16]. The ratio between the durations of these open and close phases has an impact on the harmonics and determines the spectral slope of the resulting speech signal. A long open cycle and a long closing cycle, for example, result in a sine-like wave with weak harmonics. Conversely, when the open phase is shorter, glottal airflow builds up to pulsating waves with rich harmonics [Hon08].

The fundamental frequency, denoted F_0 , is defined as the lowest harmonic measured in Hertz (Hz), which corresponds to cycles per second of the vocal folds. Its regulation is influenced by the stiffness and length of the vocal folds, which in turn are affected by the position and movements of the surrounding muscles and cartilages. Aerodynamic conditions arising from the subglottal pressure have also been shown to impact F_0 [Hon08]. The fundamental frequency is also often referred to as pitch, which is more related to the subjective perception of speech on a musical note, whether it is perceived as high or low [Hir21].

After passing through the glottis, the airflow is modulated by the vocal tract, which acts as an acoustic filter to produce vowel sounds. Articulatory features such as the position of the tongue and the shape of the oral cavity cause multiple reflections within the vocal tract, giving rise to characteristic formant frequencies – primarily F_1 , F_2 , and F_3 – that distinguish different vowels [Hon08].

As the oscillation of the vocal folds during natural speech is not exactly periodic, deviations in the fundamental frequency are inevitable. These variations are referred to as jitter for perturbation in frequency and shimmer for perturbation in amplitude. They appear within a limited range of a few percent in all normal voices [Hon08].

Another crucial determinant of voice quality is noise, which refers to the amount of inharmonic energy present in the speech signal. More specifically, it is commonly quantified as the ratio between harmonic and noise components of the spectrum, often termed the harmonics-to-noise ratio (HNR) or its inverse, the noise-to-harmonics ratio (NHR). Noise in speech arises primarily from turbulence generated by airflow interruptions at and below the glottis [Zha16]. Research indicates that the perception of harmonic and noise parameters is not independent but rather involves complex interactions – their relative levels, spectral shape, and context influence how noise is perceived [Kre12].

In conclusion, the production of voice is primarily determined by adjustments in pitch, loudness, and quality of voice, which collectively serve as the source of phonation. Subsequently, this source is modulated by articulatory organs, including the lips and tongue, to shape speech sounds [Hon08]. Furthermore, these modulations give rise to prosodic patterns, encompassing syllabic emphasis, intonation contours, phrase boundaries, and sentence modality [San08].

2.2.2 Pathology

The initiation of speech generation does not occur on the physiological level as outlined in the previous chapter but rather starts with the formation of the linguistic form in our brain. This process is also known as the speech chain [Den01]. Any impairment to this chain, attributable to either cognitive or physiological factors, has the potential to adversely influence voice and speech, thereby impacting an individual's communication capabilities. This domain of study is commonly referred to as speech pathology, encompassing various groups, including fluency disorders (i.e. stuttering), speech disorders (neurogenic or structural), and voice disorders [Ken98]. The latter is typically characterized by abnormalities in pitch, loudness, and quality of voice. These impairments can be attributed to various factors, including vocal misuse and hyperfunction, leading to physical changes in the vocal folds, or they may be indicative of underlying medical or psychological conditions [Ram98].

Neurological voice disorders are characterized by impaired neuromuscular function of the larynx, with many of these diseases originating in the central nervous system. For instance, lesions in the brain and its blood vessels, consequent to a stroke, can result in impaired articulation and quality, depending on the location of the event [Wan22]. Another example is Parkinson's disease, which includes, among others, motor problems with deficits in scaling movement force and amplitude, as well as sensory problems, resulting in a distorted self-perception of voice. This lack of muscle coordination leads to characteristics such as monotonous pitch, reduced vocal loudness, as well as imprecise prosodic articulations like short rushes of words and dysfluency [Ram11].

Functional dysphonia, another category of voice disorder, is primarily caused by psychological conditions. Emotional stressors such as anxiety and depression can lead to a hypercontraction of the laryngeal muscles [Roy03]. Furthermore, since depression also reduces cognitive abilities, actions such as speech planning and motor coordination processes are affected. Consequently, depression often leads to a slower speech rate and decreased pitch variability [Cum15].

Given the prevalence of diseases affecting speech production, the analysis of voice characteristics provides a non-invasive, cost-effective, and privacy-preserving method for detecting and monitoring such conditions.

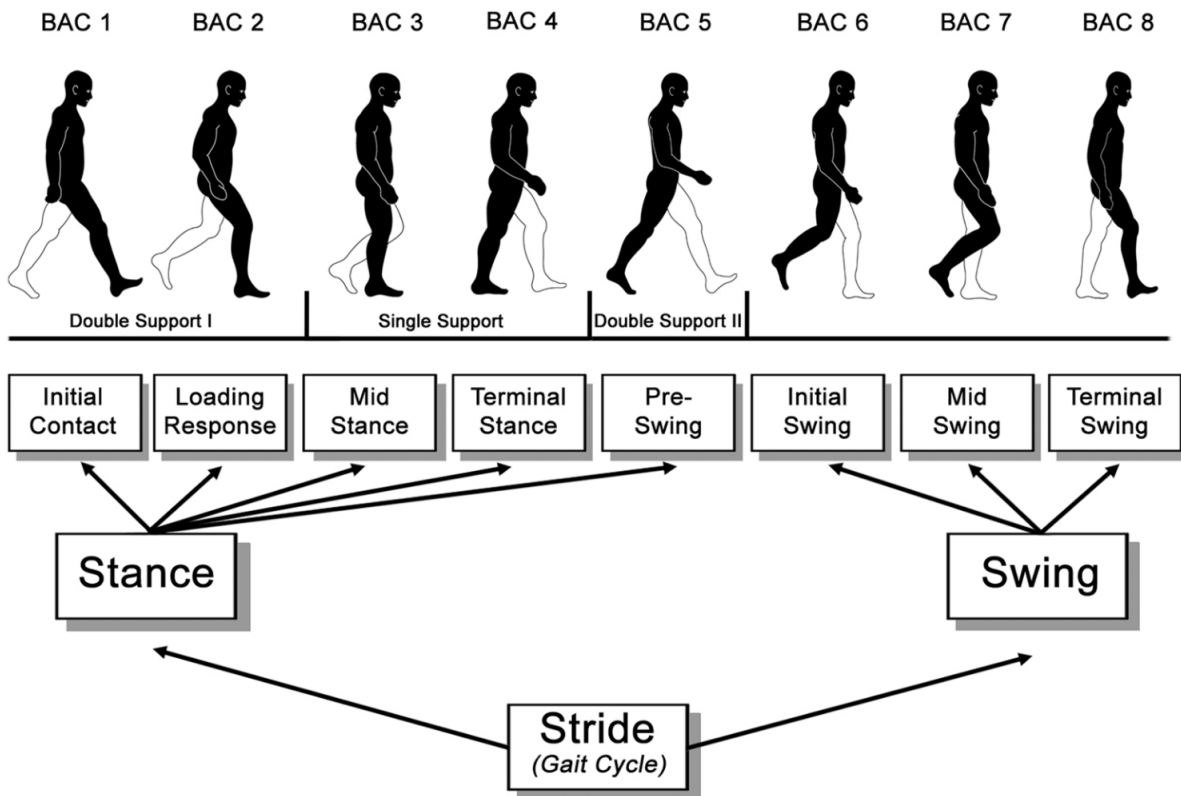


Figure 2.2: Functional divisions of the gait cycle according to [Per10], Figure adapted from [Stö15].

2.3 Gait and Motion

2.3.1 Walking Patterns

The act of walking is a complex task that involves a multitude of mechanical and neuromuscular processes within the body [Cha02]. Human locomotion is characterized by a repetitive, oscillatory sequence of limb motion, which is employed to propel the body forward while also maintaining stance stability [Tag95]. This sequence can be delineated as a gait cycle, as illustrated in Figure 2.2.

Humans move forward by shifting one limb into a new position, while the entire weight of the body remains on the other limb and vice versa. A complete cycle, which is also referred to as stride, is complete when the foot on the same side is again in contact with the floor. The two periods of the gait cycle are termed stance and swing. The stance phase is defined as the entire time that the foot is in contact with the ground, beginning with the initial contact and ending with toe-off. This phase constitutes approximately 62 % of the entire cycle and is succeeded by the swing phase, which commences with the elevation of the foot from the ground until it once again makes contact with the floor. This phase persists for the remaining 38 % of the cycle and can be

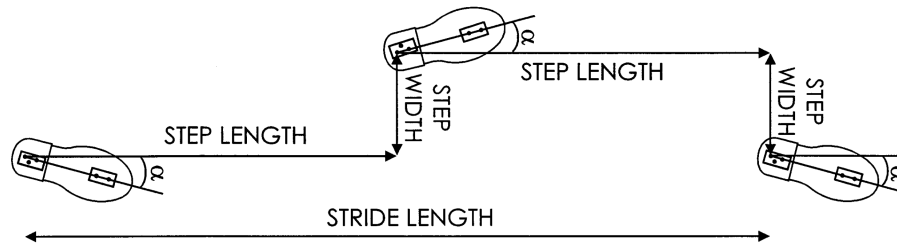


Figure 2.3: Illustration of measures within a gait cycle: stride and step length, step width, and foot angle. Figure adapted from [Sto98].

further subdivided into initial, mid, and terminal swing, depending on the advancement of the limb. Double support time involves the weight transfer from one limb to the other, concurrently maintaining forward progression in walking. It is a critical moment for stabilizing the body during gait. Subsequently, during the single support phase, one limb supports the entire weight of the body until the terminal stance [Per10].

In addition to stance, swing, double support, and cycle duration, gait is also characterized by other measures. Stride and step length, as well as step width and foot angle, are also important dimensions in gait and defined as seen in Figure 2.3. With these measures, the gait velocity and cadence, i. e. steps per minute, can be calculated [Sto98]. Regular and symmetrical cycles are indicative of normal gait, while deviations from these patterns are suggestive of neurological diseases [Moo16] and mood disorders [Fel20].

2.3.2 Motion Analysis

In order to identify any potential abnormalities in gait, balance, and posture, it is necessary first to perform a comprehensive evaluation of the movement of the person. This assessment may take various forms, ranging from simple observation to different techniques of automated motion analysis [Cha02].

Stationary settings include image-based technologies, such as motion capture systems. These consist of multiple cameras and a tracking system comprising markers attached to the body, with the purpose of generating three-dimensional data, thereby providing detailed insights into kinematics. Walkway technologies utilize pressure-sensitive plates embedded in the floor, with the function of assessing kinetic forces during movement [Cha02].

Wearable technology provides a different approach to motion analysis, requiring placement of devices on various parts of the body. This includes pressure and force sensors, inertial sensors, smartphones, and electromyography. IMUs represent a prevalent method, comprising an ac-

celerometer and a gyroscope [Mej21]. These sensors possess the capacity to measure acceleration signals, which are generated during general movement and walking, as well as the angular rate due to rotations during motion [Guj19].

Environmental technologies, including image-based and walkway systems, provide controlled setups, are non-intrusive and offer high precision. However, they are often confined to controlled laboratory environments, which may not reflect natural gait patterns [Mej21]. In contrast, wearables are also applicable in real-world environments, but ensuring correct and consistent usage remains a challenge, as user compliance and proper placement of sensors are critical for accurate data collection [Yus16].

2.3.3 Gait Analysis

To derive typical spatiotemporal gait characteristics such as step length or stride time from IMU data, several algorithmic steps are required. An overview of the process is illustrated in Figure 2.4.

Starting with the raw IMU signal, the first step involves identifying regions of interest, specifically segments likely to contain walking activity, while excluding non-walking periods such as standing still. The detected walking regions are then segmented into individual strides, in which key gait events, such as initial contact and toe-off, are identified [Küd24]. Based on these detected events, foot trajectories are reconstructed step-by-step. Further integration of the IMU signals allows estimation of the foot's position in space. To counteract errors due to velocity drift, static periods through *zero-velocity detection* are employed for correction [Kit16]. Finally, spatial (e.g., stride length) and temporal (e.g., stride time) parameters are extracted from the reconstructed trajectories [Küd24].

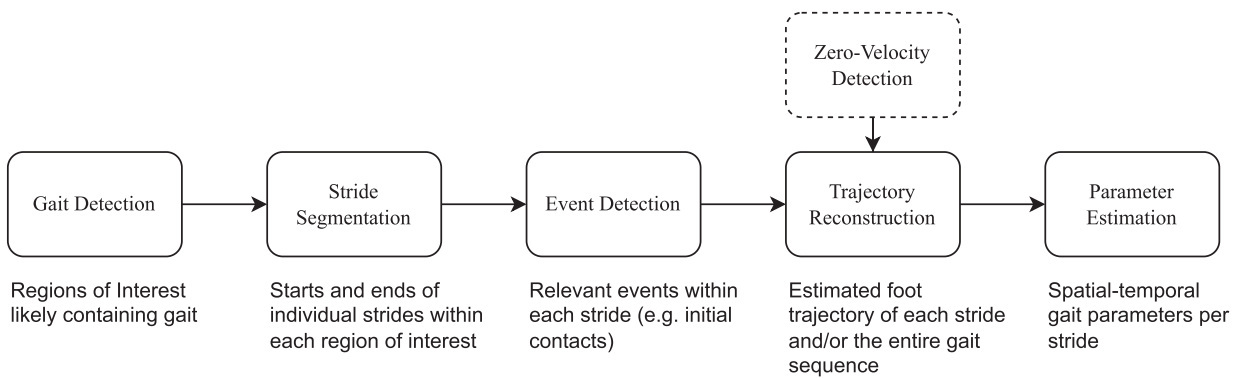


Figure 2.4: Gait analysis pipeline, adapted from [Küd24].

2.4 Machine Learning

Machine learning (ML) encompasses algorithms that enable computers to learn patterns from data and make predictions without being explicitly programmed. ML is commonly categorized into supervised, unsupervised and reinforcement learning [Uça20]. Since this thesis focuses on classification tasks, the discussion in this chapter will concentrate on supervised classification techniques.

In supervised classification, an algorithm learns the relationship between observable input features X and their corresponding target labels y . The primary objective is to generalize this learned relationship in a way that allows the model to perform accurately on previously unseen data. A variety of classifiers are available for this purpose, including kNN, SVM, LR, and deep neural networks (DNNs) [Dom12].

The general procedure for training and evaluating a ML model is illustrated in Figure 2.5. To properly assess the generalization capability of a model, the dataset must first be partitioned into distinct training and test sets. Importantly, the test set must remain strictly separated from the training process. During the training phase, the model's parameters are iteratively adjusted to minimize the difference between the predicted and the actual output, quantified by a loss function. Techniques such as cross-validation (CV) are employed during this phase to mitigate overfitting [Dom12].

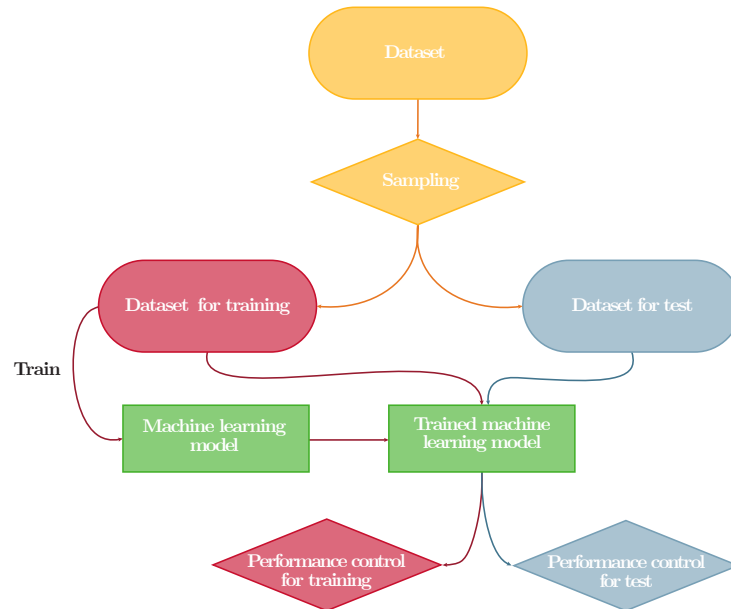


Figure 2.5: Training and testing process, adapted from [Uça20].

Once training is completed, the fully trained model is evaluated using the previously unseen test set. This evaluation phase, highlighted in blue in Figure 2.5, provides an unbiased estimate of the model's performance. Common evaluation metrics include accuracy, sensitivity (true positive rate), specificity (true negative rate), as well as area under the curve (AUC), among others [Uça20].

In the following sections, the specific classifiers and models utilized in this thesis will be introduced in greater detail.

2.4.1 Logistic Regression (LR)

In logistic regression (LR), a set of independent variables X is used to predict a binary dependent variable y . This model addresses the problem of mapping continuous input variables to a dichotomous output by employing a probabilistic approach. The most commonly used function for this multivariate task is the logistic sigmoid function, defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (2.1)$$

This function maps any real-valued input to the interval $[0, 1]$, which can be interpreted as the probability that the input X belongs to a particular class, typically $y = 1$. Accordingly, the probability that a given instance x belongs to class 1 is modeled as:

$$P(y = 1|x) = \sigma(\omega^T x + b) = \frac{1}{1 + e^{-(\omega^T x + b)}} \quad (2.2)$$

Here, $\omega \in \mathbb{R}^d$ is the weight vector, and $b \in \mathbb{R}$ is the bias term. These unknown parameters are estimated from a training dataset using maximum likelihood estimation [Kle02].

2.4.2 k-Nearest Neighbors (kNN)

The k-nearest neighbors (kNN) method is a widely adopted, non-parametric technique for classification tasks [Fix51]. The algorithm assigns a new sample to the class most frequently represented among its k nearest samples in the dataset.

For a data point from the test set, the algorithm calculates the distance to all instances in the training dataset, selects the k closest samples and determines the output label by majority vote [Pet09]. Formally, given a training set

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} \quad (2.3)$$

where $x_i \in \mathbb{R}^d$ are feature vectors and y_i are the corresponding labels, and a new instance x , kNN operates as follows:

1. Compute the distance $d(x_i, x)$ for each training instance.
2. Identify the k training instances with the smallest distance to x .
3. Set the predicted class of test sample x equal to the most frequent class among the k nearest training samples.

To calculate the distance, various metrics can be applied, such as the Manhattan or Euclidean distances. These can be used with either uniform or distance-weighted voting schemes. A commonly used distance metric is the Euclidean distance, defined as:

$$d(x_i, x) = \sqrt{\sum_{i=1}^d (x_i - x)^2} \quad (2.4)$$

This metric measures the straight-line distance between two points in a Euclidean space and is sensitive to the scale of features, which often necessitates normalization or standardization of the data.

2.4.3 Random Forest (RF)

Random forest (RF) is a widely used and versatile ensemble learning method, particularly effective for classification tasks. It operates by constructing a large number of decision trees during training and combining their outputs to form a single, aggregated prediction. This ensemble approach helps to improve predictive accuracy and model robustness.

Introduced by Breiman [Bre01], RFs consist of multiple randomized decision trees. Each tree is built using a variation of the training data and feature set, ensuring diversity among the individual learners. The process of creating multiple models on different bootstrap samples and aggregating their predictions is known as bagging (bootstrap aggregating), which is illustrated in Figure 2.6.

For each tree, a random sample $\mathcal{D}^*$ is drawn from the original set $\mathcal{D}$ (see Equation 2.3) with replacement (*bootstrapping*), meaning some samples may appear multiple times while others may be omitted. During the construction of each tree, at every split node, only a randomly selected subset of features is considered. The optimal split is determined based on a chosen impurity measure, such as Gini impurity or entropy. Tree construction continues recursively, splitting nodes

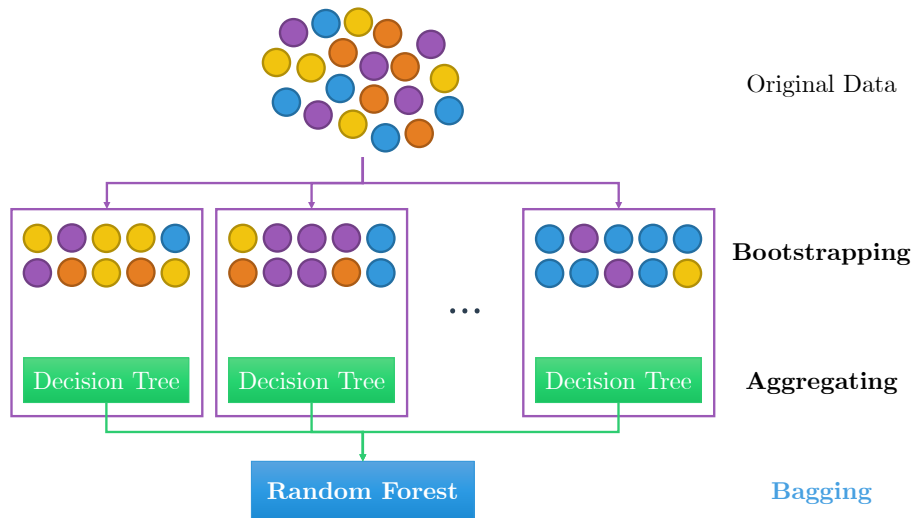


Figure 2.6: Sampling principle of a random forest. Adapted from [Sir20].

until a stopping criterion is met (e. g., a minimum number of samples required to split a node or a predefined maximum tree depth) [Bia16].

To classify a new instance, each tree in the forest provides a class prediction, and the final classification is determined by a majority vote across all trees (*aggregating*).

Due to their ability to combine a large number of weak learners (i.e. individual decision trees) using the bagging principle, RFs offer high accuracy, resistance to overfitting, and robustness across a wide range of ML tasks [Bre01].

2.4.4 Support Vector Machine (SVM)

Support vector machines (SVMs) are a family of ML algorithms for learning two-domain discriminant functions. The SVM can be divided into linear and nonlinear models. The goal is to find an optimal hyperplane that will effectively separate data points from two distinct classes in a high-dimensional feature space, with the goal of maximizing the margin between these classes [Mam09].

To understand the concept, a two-dimensional feature space with linearly separable classes can be assumed. The hyperplanes are denoted by the following Equation:

$$w \cdot x + b = 0 \quad (2.5)$$

Here, w refers to the weights and b to the bias, while x denotes the input vector of the decision boundary. The classification is performed based on the sign of $w \cdot x + b$, which determines the

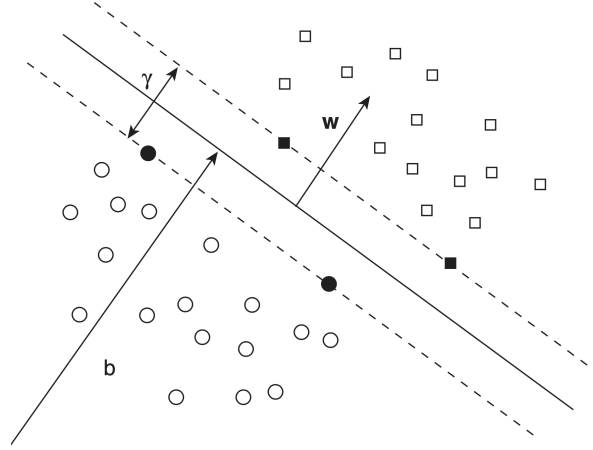


Figure 2.7: A separating linear hyperplane with margin γ for data from two classes (squares, circles), adapted from [Mam09].

predicted class label $y_i \in \{-1, 1\}$. To maximize the margin γ between the classes, the two-norm of w has to be minimized. A graphical representation can be seen in Figure 2.7.

To enhance flexibility in real-world scenarios where perfect separation is not possible, a regularization parameter C is introduced. This parameter controls the trade-off between maximizing the margin and allowing misclassifications, a technique known as the soft margin. Additionally, for datasets that are not linearly separable, kernel functions can be employed to map the data into a higher feature space, where a linear boundary can effectively separate the classes [Bis06].

2.4.5 Extreme Gradient Boosting (XGBoost)

Extreme gradient boosting (XGBoost) is a scalable and highly effective ML method, introduced by Chen and Guestrin [Che16]. Due to its strong predictive performance and computational efficiency, it has become one of the most widely adopted algorithms in ML competitions. XGBoost constructs an ensemble of decision trees, where each subsequent tree aims to correct the prediction errors made by the combined ensemble of previously built trees.

The underlying framework for this approach is known as gradient boosting, which was introduced by Friedman [Fri01]. In this framework, the model is built sequentially to minimize the loss by reducing the combined residuals of the previous trees. In contrast to models such as RFs, which employ bagging to train trees independently, boosting methods build trees iteratively, with each tree attempting to incrementally reduce the total model error.

XGBoost improves upon standard gradient boosting by incorporating a regularized objective function, which explicitly penalizes model complexity to mitigate overfitting. The best split at

each node is determined by maximizing the gain of a split, which quantifies the improvement in the objective function resulting from the split. The gain is defined as:

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} + \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (2.6)$$

where λ is the regularization parameter and G and H are the accumulated first and second-order gradients for each of the left and right child nodes, respectively. If the calculated gain is smaller than γ , the split is not performed, and the node becomes a leaf. This mechanism allows XGBoost to perform tree pruning, which helps to prevent overfitting by limiting unnecessary complexity [Che16].

2.4.6 Transformer

The transformer is a deep learning (DL) architecture based on an attention mechanism proposed by Vaswani et al. [Vas17]. Initially developed for natural language processing, it has since been employed in other domains, including computer vision [Bi21] and audio [Bae20].

As illustrated in Figure 2.8, the Transformer comprises N encoder layers (left) and N decoder layers (right). The encoder maps an input sequence x to a continuous representation sequence z . This representation is subsequently passed to the decoder, which generates the output sequence y .

The input tokens are first mapped to vector representations through an embedding matrix. To provide the model with positional information, sine and cosine functions are applied as positional encodings. This step is essential because the transformer has no inherent sequence awareness. Each encoder layer includes a multi-head attention mechanism that performs self-attention, enabling the model to capture contextual information by relating different positions within the input sequence. Each attention layer is followed by a residual connection and layer normalization. To facilitate auto-regressive generation, the decoder uses the previously generated output as additional input when generating the next. To prevent the model from accessing future tokens, a mask is applied to the first attention block. Additionally, the decoder incorporates a second multi-head attention block that performs cross-attention by integrating the encoder's output with the decoder's current state. Both the encoder and decoder contain a fully connected, position-wise feed-forward network at the end. Finally, a linear layer followed by a softmax function is used to predict the probability distribution for the next token.

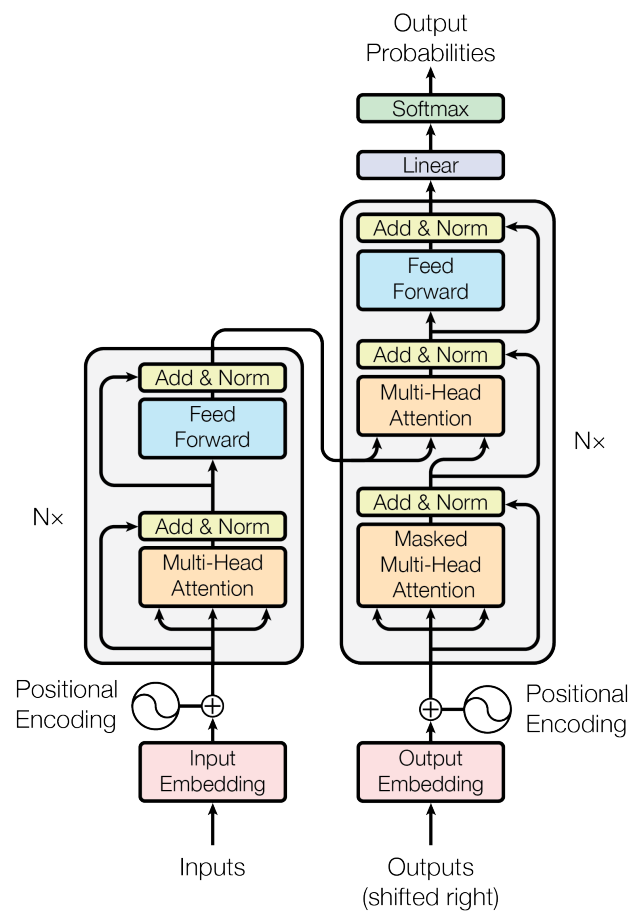


Figure 2.8: The Transformer model architecture, adapted from [Vas17].

Chapter 3

Related Work

3.1 Voice-Based Depression Detection

Individuals diagnosed with depression often exhibit distinctive speech characteristics that diverge from those observed in healthy individuals. These differentiating properties were used in early research for discriminant analysis. Specifically, features such as fundamental frequency and formants have been identified as effective markers for detecting depression [Fra00]. With advancements in acoustic signal processing and ML technologies, these speech patterns have been further explored in the field of speech depression recognition (SDR) to enhance both the detection and severity estimation of depression [Yam20].

Feature-based Machine Learning Approaches

Early research in SDR primarily employed feature-based classification and regression algorithms, including SVM [Val16; Jia18; Nas16], gaussian mixture model (GMM) [Hel13; Par17], RF [Esp21], and multi-layer perceptron (MLP) networks [Alg13; Shi21]. These models mainly use hand-crafted features extracted from speech signals, including mel frequency cepstral coefficients (MFCC), as well as spectral and prosodic features.

Numerous studies have demonstrated the effectiveness of these ML methods. For instance, as part of the Audio/Visual Emotion Challenge (AVEC) 2016, a regularly organized event, an SVM-based model was presented for depression classification and achieved an F1 score of 0.410 as a baseline for the launched challenge [Val16]. In this context and beyond of it, various other studies have been carried out on the subject of depression recognition and demonstrated the potential of feature-based classifiers methods, including SVM and GMM as well as MLP, RF,

and LR models [Hel13; Alg13; Nas16; Par17; Jia18; Shi21; Esp21]. These approaches reported different accuracy rates, ranging from approximately 60 % [Shi21] up to 90 % [Par17]. The observed differences in performance can be attributed to the selected feature sets and classifier configurations. While some studies employed comprehensive combinations of prosodic, glottal, and spectral features [Alg13; Par17; Shi21; Esp21], others focused on specific frequency-related feature sets, such as formant frequencies [Hel13] or MFCC alone [Nas16]. The latter approaches may struggle to capture the complex speech patterns associated with depression. A notable contribution in this context was the study of Jiang et al. [Jia18], which highlighted the importance of feature selection tailored to individual classifiers. Their investigation demonstrated that the optimal feature set varied across models. Their best performing method incorporated an ensemble logistic regression model alongside with a combination of different feature sets, outperforming other classification methods focusing on singular feature sets. A further challenge identified in the literature is the predominance of English-language data in SDR research [Hel13; Alg13; Nas16; Par17]. While some studies have examined non-English speech data, including Chinese [Jia18], Korean [Shi21], or Portuguese [Esp21], these efforts often faced limitations such as small sample sizes [Shi21; Esp21], unbalanced group sizes [Esp21], or the effort to label speech segments manually [Jia18].

Comprehensive review articles provide further insights into these methods, offering comparative analyses of different feature categories and classifier performances [Low20; Wu23a]. These reviews emphasize key challenges, notably the need for more diverse datasets to improve model generalization and enhance adaptability across linguistic and cultural contexts.

In conclusion, feature-based ML models established a solid foundation for voice-based depression detection, achieving notable success when appropriate features and classifiers were employed. However, limitations in generalization, especially regarding language diversity and cross-domain adaptability, have motivated researchers to explore alternative methods.

Deep Learning Approaches

With advancements in DL, the focus has increasingly shifted toward neural network-based architectures for automatic feature learning and classification. In particular, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have demonstrated promising results in this domain. While these models are designed for automatic feature extraction, some approaches still incorporate hand-crafted audio features for both classification [Tia23] and regression tasks [Yan17]. In these cases, low level descriptors (LLD) have been employed as audio features within CNN architectures.

Alternative methods have explored the potential of capturing temporal and spatial relationships within the frequency domain. For instance, long short-term memory (LSTM) networks have been combined with MFCC to exploit temporal dynamics in speech signals [Rej22]. Similarly, graphical representations of speech spectrograms have been processed using a CNN [Váz20], sometimes enhanced with residual neural network (ResNet) blocks [Ch119; Sup23]. These approaches reported accuracies between 74 % and 87 %. In other attempts, some researchers have adopted multi-network architectures that combine various feature representations. For example, the EmoAudioNet [Oth21] integrates both spectrogram-based and MFCC-based CNN models for a classification task. Similarly, He and Cao [He18] combined an LLD-based CNN with a spectrogram-based CNN to predict BDI-II scores. These hybrid approaches resulted in an accuracy of 73 % and a root mean square error (RMSE) of 9.9, respectively.

Despite the potential of DL models, their performance heavily depends on the availability of large, diverse datasets. Compared to feature-based methods, DL approaches typically require more data to train end-to-end models. However, acquiring such data can be challenging due to privacy concerns and the extensive manual effort required for labeling sparse data [Wu23a]. To address this limitation, recent approaches have leveraged self-supervised learning techniques. For instance, Zhang et al. [Zha21] proposed DEPA, an audio embedding pretrained on both in-domain and out-domain, publicly available datasets. This embedding was processed by a bi-directional LSTM, reaching a F1 score of 0.94 in classification and a RMSE of 5.73 in regression tasks.

Another strategy to mitigate the problem of data sparsity involves the use of speech foundation models, which are able to reach state-of-the-art results in SDR [Wu23b]. Han et al. [Han24] developed a model that captures deep spatial-temporal features by combining a wav2vec transformer model for feature extraction with dedicated networks for spatial and temporal information processing. Similarly, AudiBERT [Tot21], a model integrating both linguistic and paralinguistic content, has shown promising results in improving classification performance over previous models.

Review articles emphasize that DL can either be used as a classification method for hand-crafted features or as end-to-end systems that leverage deep-learned feature representations [Wu23a; Lea24]. However, these reviews also highlight persistent challenges in SDR, notably the limited availability of labeled datasets. Data sparsity arises from the fact that labels are often assigned to entire interviews rather than individual speech segments, limiting the granularity of training data. Additionally, foundation models that analyze speech content raise potential privacy concerns, further complicating data acquisition.

In summary, DL methods have expanded the scope of SDR research by offering powerful tools for automatic feature learning and improved classification performance. However, persistent

challenges related to data availability, labeling efforts, and privacy considerations continue to influence the future direction of research in this field.

Feature Extraction and Datasets

A wide range of speech feature extraction methods have been employed in SDR. Among these, MFCC remain a prominent choice in both classical ML and DL approaches due to their effectiveness in capturing essential spectral characteristics of speech signals [Nas16; Oth21; Rej22]. Another widely adopted method is the extraction of LLD from the Geneva Minimalistic Acoustic Parameter Set (GeMAPS), which is specifically designed for affective computing and psychological assessment tasks [Eyb16]. The openSMILE toolkit [Eyb10] is frequently utilized for extracting GeMAPS features, particularly in studies investigating prosodic and phonetic speech characteristics across both the ML [Val16; Yan17; He18; Shi21] and DL models [Yan17; Tia23].

In addition to feature extraction techniques, several benchmark datasets have played a critical role in SDR research, such as the AVEC dataset, which has been utilized in various studies [Val16; He18]. Another widely adopted dataset is the Distress Analysis Interview Corpus (DAIC), which has also been referenced in multiple studies [Tia23; Rej22; Han24]. The latter is a multi-modal collection of semi-structured clinical interviews, conducted both through face-to-face interactions and a virtual animated interviewer. This dataset includes several hundred interviews, with coding for depression, post-traumatic stress disorder, and anxiety [Gra14]. Importantly, it is publicly available for research purposes.

However, certain limitations must be acknowledged. The DAIC dataset, for example, is composed exclusively of English-speaking participants and features audio recorded using lapel microphones, which may restrict the generalizability of trained models to other languages and recording conditions. To address these limitations, some researchers have focused on collecting speech data from diverse linguistic and cultural backgrounds. These efforts include exploratory studies with small sample sizes [Esp21] as well as large-scale data collection initiatives [Yam20].

3.2 Gait and Motion Analysis in Depression Detection

Motor behaviour and motor regulation are intrinsically linked to emotional responses, and consequently, play a role in mental illness [Fel20]. Psychomotor changes manifest across multiple motor domains, including facial expressions, gesticulations, locomotion, and fine motor tasks [Sob97].

Research has identified significant differences in spatiotemporal gait parameters, laying the foundation for the development of machine learning methods for depression detection. One of

the early studies in this area was conducted by Lemke et al. [Lem00], who performed statistical analyses comparing gait patterns between depressed and healthy individuals. The results showed that depressed patients exhibited significantly lower gait velocity, reduced stride length, longer double limb support time, and extended cycle duration compared to the control group. Radovanovic et al. [Rad14] later supported these findings in their study and concluded that depressed individuals have less stable gait patterns and prioritize stability over speed.

Camera-Based Approach

One common method for depression classification is the employment of Kinect cameras, which facilitate the recording of an individual's gait and posture, and further extract joint positions. Li et al. [Li21], for instance, recorded the movement data from both depressed and non-depressed individuals using Kinect cameras. Utilizing the three-dimensional data captured, the researchers achieved an optimal classification outcome employing a gradient boosting algorithm, attaining an accuracy of 77 % surpassing the accuracies of SVM, LR, and RF. Similarly, Zhao et al. [Zha19] utilised Kinect cameras to predict the mental state of individuals, focusing on the individual items of the PHQ-8. Their analysis revealed significant disparities in the detection of individual symptoms based on gait measures.

Other researchers have explored ways to refine feature extraction for improved classification accuracy. Wang et al. [Wan21] used the data obtained by the Kinect camera to extract meaningful feature spaces, including spatio-temporal features, such as body sway, arm swing, walking speed, and time-domain as well as frequency-domain features. The SVM-based model, tested on various feature set combinations, achieved a result of 0.93 AUC, in which time-domain and frequency-domain features contributed the most. Lu et al. [Lu21] incorporated energetic information through the introduction of a novel rigid-body representation to reduce redundancies in raw three-dimensional data. This not only improved processing efficiency, but also boosted SVM classification accuracy to 0.89, with substantial improvements in specificity.

Deep learning approaches have also been explored. Shao et al. [Sha22], for example, analyzed the skeletal data obtained by Kinect devices with a LSTM-network. Additionally, they extracted the silhouette of the recorded individuals and further fed this data into a CNN. The fusion of these results yielded the highest accuracy of 85 %, accompanied by an enhancement in the model's generalization capability.

While camera-based approaches have demonstrated considerable promise, their practical implementation remains largely confined to laboratory settings. Furthermore, these methods often involve highly personal data, including full-body and facial recordings, raising concerns regarding

privacy and data protection. In response to these challenges, researchers continue to explore alternative motion capture techniques, which offer practical deployment in real-world scenarios.

Sensing with Wearables

To overcome the limitations associated with camera-based systems, wearable sensor technology offers a passive and continuous approach to motion analysis, enabling depression detection in real-world environments. For instance, O'Brien et al. [OBr17] recorded accelerometer data from a device worn on the wrist over the course of several days and found that depressed individuals exhibited lower activity levels, reduced entropy, and less jerk in movement, from which the researchers concluded that this indicated slower fine movements.

Jung et al. [Jun22] equipped 45 elderly individuals with an IMU sensor, which was positioned on the lower back. Participants were then observed as they walked on a 7-meter path, while several parameters of gait acceleration were measured. The classification network, a bidirectional LSTM, was able to distinguish between individuals who were non-depressed and those suffering from major depression with 96 % accuracy. In a similar study, Boolani et al. [Boo23] conducted a balance assessment as well as a two-minute walk test on a 6-meter track. They then proceeded to categorize the participants into three distinct groups, namely non-depressed, mild depression, and moderate-high depression. The achieved accuracy rate in identifying individuals with moderate-high depression compared to those not depressed was found to be 85 %, utilizing a gradient boosting model.

A number of other studies expanded the scope to a wider range of passive sensing data. Narziev et al. [Nar20] collected data from smartphones and smartwatches over a period of four weeks. Among their predefined five symptom clusters – including mood, social activity, and food intake – physical activity, primarily measured through accelerometer and step data, was the highest contributor to depression classification accuracy after sleep. Their RF classifier demonstrated an accuracy exceeding 90 %, outperforming SVM. In a similar study, a smartphone and a wristband were utilized to collect multi-modal data with the objective of predicting HAM-D scores [Gha17]. In this study, lower motion levels and reduced motion variability were also identified as key indicators of depression. They attained a low RMSE of 4.5, using a custom ensemble model.

Although wearable sensors provide a more discreet and flexible alternative to camera-based motion analysis, their reliance on additional devices worn by individuals may still pose practical challenges in real-world applications. This requirement can limit widespread adoption, particularly in scenarios where continuous monitoring without user intervention is essential.

3.3 Fusion

Given the individual strengths and limitations of speech, gait, and motion analysis for depression detection, recent research has increasingly explored the fusion of multiple modalities. For example, the combination of audio with visual cues – such as facial landmarks – is a common approach for detecting depression. Different studies employed a logical AND-function to simply connect the results of two different classifiers, independently analyzing audio and video [Val16; Nas16]. Expanding on this, Yang et al. [Yan17] additionally introduced a text-based classifier, yet used a DNN to integrate results from audio, video, and text, thereby enhancing the overall detection performance.

Section 3.2 already discussed multi-modal passive sensing. In this context, different fusion strategies have been explored. Gruenerbl et al. [Gru14] combined smartphone-derived GPS and accelerometer data using a weighted fusion process to enhance depression detection accuracy, while Ghandeharioun et al. [Gha17] and Rykov et al. [Ryk21] fused the most significant biomarkers for their depression detection model on the feature level. They extracted meaningful features from wearables and smartphones, including sleep patterns, motion, communication frequency, location, and physiology signals.

More recent studies also investigated the fusion of speech and gait data. Dai et al. [Dai23] applied a linear weighted score of video, audio, and gait data extracted from video recordings, leading to notable improvements in classification performance. Li et al. [Li25] combined audio and video data, from which they also extracted head movements, during an interview. Additionally, they incorporated gait data from an IMU worn on the ankle, as their study focused on depression detection in patients with Parkinson’s disease. By integrating these modalities, their model achieved an accuracy of 95 %, representing a more than 10 % improvement compared to their uni-modal audio-based model.

3.4 Ear-Worn Sensors

The utilization of IMU sensors, inherently integrated in wearable devices, has become increasingly prominent for healthcare monitoring due to their compact size. While these sensors are commonly positioned on the wrist, lower back, or ankle, they can be seen as quite obtrusive as they need to be worn additionally, which can limit the wearing comfort [Día20]. For a more seamless integration into everyday life, an emerging alternative involves the use of ear-worn sensors, also referred to as earables. These devices offer a discrete and convenient means of capturing a variety of vital signals and health-related bodily functions [Röd22].

Earables have demonstrated capabilities in recording electroencephalogram (EEG) and photoplethysmogram (PPG) data, both of which provide valuable insights into mental well-being. Athavipach et al. [Ath19] and Li et al. [Li17] successfully developed an in-ear device to record EEG signals for emotion recognition. Both studies employed an SVM-classifier, with Athavipach et al. reporting an accuracy comparable to other head-worn devices, achieving 71 % accuracy in valence and 73 % in arousal classification. A separate study utilized an ear-worn optical sensor to obtain PPG data, which they combined with IMU data to classify individuals as either stressed or non-stressed [Bar23]. Their CNN-based approach achieved an accuracy of 92 %.

Additionally, earables have also been proven effective in tracking behavioral and facial movements using IMU sensor data and different machine learning techniques. These approaches achieved high accuracy across tasks like bruxism detection [Bon21], facial expression recognition [Ver21], and head and mouth movement classification [Hos19]. With models like kNN, RF, SVM, and neural networks they reached accuracies of up to 90 %.

Ear-worn IMU sensors have further been applied to fitness and activity tracking. These approaches achieved high accuracy in tasks such as step counting [Pra20], energy expenditure estimation [Bou14], and daily activity classification [Bou14], reaching up to 98 % across different feature-based ML models and deep neural networks. Some studies additionally incorporated audio from the earbud's microphone to identify activities like conversation and head gestures [Min18], achieving average accuracies of around 88 %.

Earables have also shown promise in the emerging field of gait analysis. Researchers have validated the use of commercially available earbuds for detecting gait parameters during walking and running, using reference systems for comparison [Jun23]. Other studies explored the potential of ear-worn IMU sensors to estimate key metrics in both normal and pathological gait patterns [Dia20], as well as to detect specific gait events such as heel strike and toe-off [Hwa18]. In addition, head-worn approaches have demonstrated potential for real-world monitoring and real-time feedback, a benefit particularly for individuals undergoing rehabilitation due to injury or aging, further suggesting the incorporation of gait analysis into consumer ear-worn devices [Hwa21].

Given the established links between depression and altered physiological, behavioral, and activity patterns, earables represent a promising instrument for passive depression detection, seamlessly integrated into everyday devices such as earbuds and HA. Their unobtrusive appearance, in addition to their ability to monitor various modalities, offers a viable approach for continuous mental health assessment.

Chapter 4

Data Collection

4.1 Study Design

As, to the best of our knowledge, no publicly available dataset featuring ear-worn IMU exists, a study to collect original data was conducted. Given the limited prior research in this area, an exploratory study was designed. Therefore, the study involved only a modest sample size, comprising an equal number of participants from a healthy control group and a group experiencing symptoms of depression.

The study was conducted by the Machine Learning and Data Analytics Lab at Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany, in collaboration with the Department of Clinical Psychology and Psychotherapy at FAU, ensuring the integration of clinical expertise in working with individuals affected by depression. The study was approved by the local ethics committee at FAU (Re-No. 25-22-S), and all participants provided written informed consent prior to participation. The study protocol can be found in Appendix A.

Recruitment and Screening

Participants were recruited via email newsletters, online forums, public bulletin boards at the university, and flyers posted in doctors' offices in the city of Erlangen.

Given the linguistic variability in speech features [Sch16], only native German speakers were eligible for participation. Additional inclusion and exclusion criteria (detailed in Table 4.1) were evaluated using an online screening questionnaire. This questionnaire also included the PHQ-8, a validated screening tool for depression. Participants scoring ≥ 10 in the PHQ-8 were classified as experiencing depressive symptoms; those scoring < 10 were considered healthy. Furthermore, it

Inclusion Criteria	Exclusion Criteria
• Age ≥ 18 years • German speakers • Ability to walk independently • Willingness and ability to participate	• Psychotic or bipolar disorder • Acute addiction • Neurodegenerative or musculoskeletal diseases; impairment of mobility • Non-consenting persons • Supported hearing loss (wearing a HA in everyday life)

Table 4.1: Inclusion and exclusion criteria for study participation.

was determined whether psychological self-perception corresponds to the result of the PHQ-8.

Only individuals whose responses met all eligibility criteria were invited to take part in the on-site study. All sessions took place at the Machine Learning and Data Analytics Lab at FAU, with each session lasting approximately one hour. Participants were briefed on the procedure, completed the PHQ-8 again to obtain an up-to-date score, and provided sociodemographic information. To ensure privacy, each participant was assigned a pseudonymous code for data management.

Study Population

In total, 48 individuals completed the online screening. Following exclusion based on the predefined criteria, 25 participants were invited to the lab. Of these, 13 were initially classified as healthy, and 12 as experiencing depressive symptoms.

Due to technical malfunctions, the data from one healthy participant were incomplete and thus excluded entirely from further analysis. Additionally, as the PHQ-8 score was determined both during the initial screening and immediately prior to the on-site study, discrepancies in classification were observed. Specifically, three of the twelve initially identified depressed participants scored ≥ 10 during screening but fell below the threshold at the time of the study. Their status is considered as indeterminate, which is why they are excluded from subsequent analyses unless explicitly stated otherwise. This resulted in a sample of twelve healthy and nine depressed participants.

In total, there were 16 female and nine male participants, aged between 21 and 72 years. The healthy control group consists of eight female and four male participants, with a mean age of 39.8 years. The depressed group consists of an equal gender distribution of six females and three males, with an average age of 32.8 years. An overview of the sociodemographic data is given in Table 4.2.

	Gender	n	Age	Height [cm]	Weight [kg]	Glasses [%]
Healthy	female	8	41.8 ± 20.6	163.0 ± 7.2	58.5 ± 6.4	62.5
	male	4	36.0 ± 22.1	184.5 ± 7.5	76.2 ± 6.6	75.0
	all	12	39.8 ± 20.2	170.2 ± 12.6	64.4 ± 10.7	66.7
Depressed	female	6	33.4 ± 15.8	165.5 ± 4.9	63.7 ± 12.8	50.0
	male	3	31.3 ± 11.0	180.8 ± 11.4	88.7 ± 17.0	33.3
	all	9	32.8 ± 13.7	170.3 ± 10.0	72.0 ± 18.2	44.4
All	all	25	37.6 ± 17.6	170.4 ± 11.2	69.0 ± 15.7	56.0

Table 4.2: Sociodemographic data of the participants, further subdivided by gender. Age, height and weight is specified with mean $\pm$ standard deviation. Difference to group “All” due to the four exclusions.

4.2 Procedure

The study comprised two main components: a semi-structured interview and a gait recording session. The order of these tasks was randomized for each participant to control for order effects.

During the interview, both head movement data and voice recordings were collected. Due to technical limitations, these modalities were recorded separately. Participants were equipped with a pair of Signia receiver in the canal (RIC) hearing aid dummies for voice recording, connected to an amplifier and a Tascam audio recording device. Head movements were recorded using a RIC HA of model type Signia Pure 312 with a built-in IMU of type BMI270 (3D accelerometer, ± 2 g; 3D gyroscope, $\pm 1000^\circ \text{s}^{-1}$; $f_s = 50$ Hz) as a special research setup, connected to a custom gait study app (version 1.2.0.268) via Bluetooth on an Android device. All equipment was provided by WS Audiology (Erlangen, Germany). The full setup during the interview is shown in Figure 4.1.

To align the three axes of the IMU sensor, the participant had to complete two calibration tasks prior to the interview. The first was standing in an upright position for 20 seconds, while the second one was a sit-to-stand task, where the participants were asked to stand up and sit down three times in a row.

The semi-structured interview was designed to explore participants’ current psychological and physical well-being, drawing on methodologies established in prior research [Mun12; Alg12; Sch13; Val13; Ger23]. The questions aimed to elicit both positive and negative emotional responses in a balanced manner, covering experiences from both the present and the past.



Figure 4.1: Experimental setup of the interview during the study (situation was recreated).

For the gait assessment, participants wore another Signia Pure 312 RIC HA with an identical BMI270 IMU connected to the gait study app. Since the IMU signal is placement-independent [Sei23], the device was placed on the right ear only.

Participants were asked to walk three laps along a predefined path in a secure outdoor area, at their own pace and in their usual manner. To promote natural behavior, participants were not observed during the walk by the study supervisors. After returning indoors, they completed a feedback form, again without supervision. This unsupervised period also allowed for passive head movement recording under naturalistic conditions.

More information regarding the study procedure including the exact questions asked can be found in Appendix B.

Chapter 5

Methods

5.1 Voice Analysis

5.1.1 Pre-Processing

Microphone Calibration

The interview was recorded using four tracks, corresponding to the microphones of the HAs, which are positioned at the front and at the rear of the device (left and right). All microphones were calibrated to ensure consistent and accurate audio capture. The calibration process involved applying a 1 kHz sine wave at a sound pressure level of 70 dB to each microphone. Under these conditions, recording level of -50 dB relative to the full scale was expected for each microphone. The actual measurements deviated slightly from this reference, necessitating gain corrections. The gain adjustments to be applied per track are summarized in Table 5.1.

Channel	Microphone	Gain
1	Left front	-1.98 dB
2	Left rear	-1.48 dB
3	Right front	-3.24 dB
4	Right rear	-3.33 dB

Table 5.1: Microphone channels and their corresponding gain corrections.

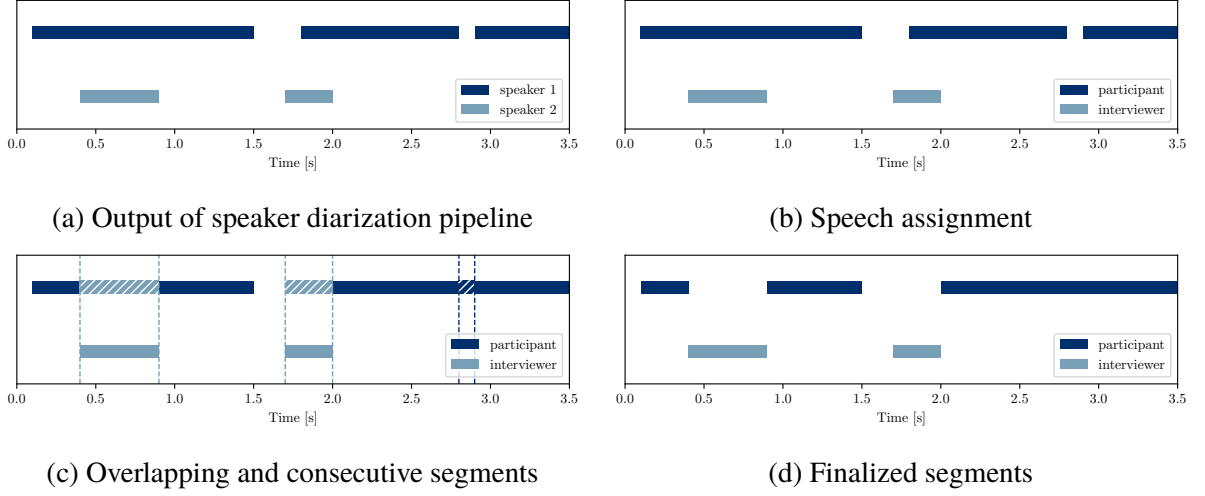


Figure 5.1: Process of getting speech segments of the participant.

Speaker Diarization

Since the audio recordings contain speech from both the participant and the interviewer, the next step involved segmenting the audio into distinct speech regions for each speaker. To achieve this, a Python-based speaker diarization model was employed [Pla23; Bre23]. The DL-based model is based on *pyannote.audio*¹ and automatically identifies the start and end points of speech segments for each speaker. These segments are visually represented in Figure 5.1a. The speaker with the majority of speaking time in total was identified as the participant (see Figure 5.1b).

To isolate only the segments where the participant speaks exclusively, the segments of the interviewer as well as the overlapping segments, where both persons speak simultaneously, were excluded. Furthermore, as shown in Figure 5.1c, consecutive speech segments from the participant, which are uninterrupted by the interviewer, were concatenated. This inclusion of pauses ensures the participant’s speech segments are properly captured for subsequent analysis. Finally, only the processed segments of the participant, marked in dark blue in Figure 5.1d, are further analyzed.

5.1.2 Feature Extraction

For the purpose of voice feature extraction, the openSMILE² toolkit [Eyb10] was employed. This toolkit enables the computation of GeMAPS, a widely adopted standardized feature set for automatic voice analysis. GeMAPS includes a range of vocal and spectral features such as the fundamental frequency F_0 , formant frequencies (F_1 , F_2 , F_3), jitter, and shimmer, as well as

¹*pyannote.audio* (Version 3.3.2), available at <https://github.com/pyannote/pyannote-audio>.

²*openSMILE* (Version 2.5.0), available at <https://github.com/audeering/opensmile>.

temporal characteristics like the mean length of unvoiced segments. In addition to these acoustic features, emotional dimensions – specifically arousal, dominance, and valence – were extracted using a pretrained wav2vec 2.0 model [Wag23].

To ensure that only the participants' speech was analyzed, features were computed individually for each of their speech segments, excluding the interviewer. For each participant, the extracted features were subsequently aggregated by computing the mean and standard deviation as well as mean and standard deviation weighted by the duration of each segment. As a result, four values were computed for each of the 88 features in the extended GeMAPS [Eyb16] and each of the three emotional dimensions, yielding a total of $(88 + 3) \times 4 = 364$ features per participant.

5.2 Gait and Motion Analysis

5.2.1 Pre-Processing

The data necessary for gait and motion analysis was obtained from different HAs containing IMU sensors across all participants. To ensure a coherent data interpretation, multiple calibration steps were applied.

Sensor Calibration

Initially, calibration was performed using the method proposed by Ferraris et al. [Fer94] to correct for sensor-specific biases and axis misalignments, enabling reliable comparisons across devices. This calibration involved placing each HA within a cube that allowed for systematic orientation. By resting the cube on its different faces, gravitational acceleration could be isolated along each axis. Furthermore, a known angular rotation was introduced by rotating the cube on a flat surface – starting and ending with one face in contact with a fixed reference – to provide reference angular velocity data. The recorded accelerometer and gyroscope data were then processed using the algorithms described by Ferraris et al. to determine necessary transformation and correction factors to mitigated systematic errors due to axis misalignment and sensor drift.

Anatomical Calibration

A second calibration step was required for data obtained under a modified experimental setup. Specifically, during interviews, audio had to be recorded using specially prepared dummies that did not allow simultaneous IMU measurements. To capture movement data, a third HA device containing an IMU was mounted on a headband worn by participants, as depicted in Figure 5.2.



Figure 5.2: Sensor position during the interview.

Due to variations in both the positioning of the headband and the sensor itself, the alignment between the sensor coordinate system and the participant's anatomical axes was unknown. To address this, anatomical calibration was performed using the approach described by Kong et al. [Kon16].

Participants were first asked to stand upright for 20 seconds, during which the vertical (z-) axis was determined from the median acceleration vector, assuming it represented gravitational acceleration. Next, participants performed three cycles of sit-to-stand movements. Principal component analysis (PCA) was applied to the resulting acceleration data to identify the dominant direction of motion, corresponding to the sagittal plane. The x-axis was then defined as the cross product of the normal vector of this plane and the z-axis. Finally, the y-axis was derived as the cross product of the z- and x-axes to ensure an orthogonal right-handed coordinate system. The transformation matrix computed from these axes was used to reorient the sensor data to align with anatomical directions.

5.2.2 Feature Extraction

IMU data were recorded at three distinct time points throughout the session to capture both gait characteristics and head movement behavior. The feature extraction process was tailored to the context of each recording.

Gait Features

During one phase of the experiment, participants walked along a predefined track while wearing the HA on their ear. The IMU embedded in the HA recorded motion data, which was subsequently

processed using *EarGait* [Sei23; Sei24], a Python-based toolbox³ specifically developed for estimating spatial and temporal gait parameters from IMU data collected at the ear. This tool automatically extracted a total of 45 gait-related features, including metrics such as step length, stride time, and other spatiotemporal parameters relevant for gait analysis.

Head Movement Features

IMU data for head movement analysis were recorded during two distinct phases, namely (1) immediately following the walking task while participants provided written feedback, and (2) throughout the interview. In the first instance, the HA remained attached at the ear, in the second, due to the use of audio-recording dummies, a separate HA was mounted on a headband.

For both recordings, a broad set of generic time-series features was extracted using *tsfresh* [Chr18], a Python library⁴ that applies a large array of time-series feature extraction methods. This includes statistical and frequency-domain features such as absolute energy, binned entropy, variance, and many others, computed across each of the six IMU data streams (i.e., three-axis accelerometer and three-axis gyroscope). To manage computational demands, especially given the long duration of the interview recordings, features with particularly high computational costs were omitted.

In addition to the generic features, expert-defined features were extracted following the methodology of Richer et al. [Ric24]. These included, for example, indicators of absolute movement and static periods, with the latter defined as time windows in which the Euclidean norm of acceleration and angular velocity remained below a predefined threshold, signifying minimal movement.

In total, 4,669 features were computed for each of the two head-movement recording segments, enabling a comprehensive characterization of participants' movement behavior during these tasks.

5.3 Evaluation

5.3.1 Statistical Analysis

For the statistical analysis, only features previously highlighted in related fields of literature were selected for further inspection. In the domain of vocal features, these included pitch, which is closely related to F_0 , shimmer, and jitter [Low20] as well as affective parameters such as arousal, dominance, and valence. For gait analysis, step length, gait velocity, and stride time were chosen

³*EarGait* (Version 3.4.0), available at <https://github.com/mad-lab-fau/eargait>.

⁴*tsfresh* (Version 0.20.3), available at <https://github.com/blue-yonder/tsfresh>.

Name	Description	Null Hypothesis	Value Interpretation
Shapiro-Wilk test	Tests whether data in a group is normally distributed.	The data are normally distributed.	$p > 0.05$: normal $p < 0.05$: not normal
Levene's test	Tests whether variances between two groups are equal.	The variances of two groups are equal.	$p > 0.05$: equal variances $p < 0.05$: unequal variances
Student's t-test	Compares means of two groups (assuming normality and equal variances).	The group means are equal.	$p > 0.05$: equal means $p < 0.05$: Significant difference
Welch's t-test	Compares means of two groups (assuming normality, but unequal variances).	The group means are equal.	$p > 0.05$: equal means $p < 0.05$: Significant difference
Mann-Whitney U Test	Non-parametric test comparing two independent groups.	The distributions are equal.	$p > 0.05$: equal distribution $p < 0.05$: Significant difference
Cohen's d	Quantifies the magnitude of the mean difference.	Not applicable.	$d < 0.2$: small, $0.2 \leq d < 0.8$: medium, $d \geq 0.8$: large effect size
Correlation r	Quantifies the magnitude of the group difference.	Not applicable.	$r < 0.1$: small, $0.1 \leq r < 0.5$: medium, $r \geq 0.5$: large effect size

Table 5.2: Overview of statistical tests and effect size measures used in the analysis.

based on their reported relevance [Lem00]. Regarding motion data, entropy and energy [OBr17], mean duration and ratio of static periods, and the mean absolute movement [Ric24] were analyzed to represent behavioral movement characteristics.

To evaluate potential group differences between participants with depressive symptoms and healthy controls, a range of statistical tests was conducted. A visual inspection of the data distributions was first performed, which included all participants, including those with an indeterminate depression status. However, for the inferential analyses, only participants meeting the defined group criteria (based on PHQ-8 scores) were included.

A summary of the statistical tests employed is provided in Table 5.2. Initially, the Shapiro–Wilk test was applied to assess the assumption of normality for all selected variables per domain within both groups [Sha65]. The null hypothesis of this test posits that the data are drawn from a normally distributed population. To examine the homogeneity of variances between groups, Levene's test

was used [Lev60], which tests the null hypothesis that the population variances are equal.

Based on the outcomes of these preliminary tests, appropriate group comparison tests were selected. If both groups exhibited normally distributed data (i. e., $p > 0.05$ in the Shapiro–Wilk test) and equal variances (i. e., $p > 0.05$ in Levene’s test), Student’s t-test was applied [Stu08]. If data were normally distributed but variances were unequal ($p < 0.05$ in Levene’s test), Welch’s t-test was used [Wel47]. In cases where normality was violated for at least one group ($p < 0.05$ in the Shapiro–Wilk test), the non-parametric Mann–Whitney U test was conducted [Man47].

To complement p-values with an indication of the practical significance of observed effects, effect sizes were calculated. For t-tests, Cohen’s d was computed to quantify the standardized mean difference between groups [Coh13]. For non-parametric tests, effect size r was determined [Cur56], defined as:

$$r = \frac{z}{\sqrt{n}} \quad (5.1)$$

where z is the standardized test statistic value from the Mann-Whitney U rank test and n is the total number of observations.

As the aim of this analysis is exploratory rather than confirmatory, no correction for multiple comparisons was applied. This decision is based on several considerations. First, the study does not seek to test predefined hypotheses for decision-making purposes, but instead to generate insights from the collected data [Ben01]. Second, corrections could increase the risk of type II errors, particularly in the context of a small sample size, potentially obscuring meaningful effects that merit further investigation [Rot90]. Nevertheless, in line with recommendations in [Was19], the results are interpreted with caution and are not overgeneralized.

5.3.2 Classification Experiments

To evaluate the potential of detecting depressive symptoms based on features derived from voice, gait, and motion data, a ML pipeline was implemented that combined various pre-processing, feature selection, and classification strategies. Each pipeline consisted of a variance threshold to remove all features with zero variance across samples. The remaining features were then normalized using either standardization (scaling to zero mean and unit variance) or min-max scaling (scaling to a $[0, 1]$ range). Dimensionality reduction was performed to mitigate overfitting and enhance generalization. Two approaches were compared, namely select k-best, which retains the top k features according to univariate statistical scores, and recursive feature elimination, using a SVM classifier as the base estimator. Classification was performed by using logistic

regression (LR), k-nearest neighbors (kNN), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGBoost).

To ensure robust model evaluation and prevent data leakage during hyperparameter optimization, all models were trained and evaluated using a stratified five-fold nested cross-validation (CV) procedure, comprising an outer loop for performance estimation and an inner loop for hyperparameter optimization.

In the outer loop, the dataset was split into five stratified folds, preserving the class distribution within each fold. For each outer iteration, one fold served as the test set, while the remaining four folds formed the training set. The test fold was exclusively used for final performance evaluation and was not involved in training or hyperparameter tuning.

Across each outer training set, another stratified five-fold CV was performed for the hyperparameter search. Within this inner CV, different hyperparameter combinations of the pipeline components (i. e., variance thresholding, scaling, feature selection, and classifier) were evaluated with the training set of each inner fold. A randomized search was used for RF and XGBoost, while a grid search was applied for all other classifiers. The optimal hyperparameter configuration was selected based on the highest average balanced accuracy across the inner folds. This configuration was then used to refit the model on the entire outer training set, which was subsequently evaluated on the outer test set. The complete range of tested hyperparameter values is listed in Appendix C.

All pre-processing steps (variance thresholding, scaling, feature selection) and model training were performed within each training fold, ensuring that no information from the outer test folds influenced the model during training or tuning.

This procedure was repeated for all combinations of pre-processing, feature selection, and classifiers. With one variance thresholding method, two scaling methods, two feature selection techniques, and five classifiers, a total of 20 distinct pipeline configurations were evaluated using the described nested CV procedure. All experiments were performed using the Python libraries *BioPsyKit*⁵ [Ric21] and *scikit-learn*⁶ [Ped11].

Model performance was assessed using balanced accuracy, precision, recall, and F1 score. To gain insight into the decision-making processes of the models, confusion matrices and SHapley Additive exPlanations (SHAP) values were used. SHAP provides a consistent and model-agnostic approach to explainability by assigning contribution values to each feature for individual predictions [Lun17]. Since the primary goal of this analysis was to examine the impact of different feature domains rather than to compare classifiers, a single representative classifier is selected for each domain to visualize and interpret results.

⁵*BioPsyKit* (Version 0.13.0), available at <https://github.com/mad-lab-fau/BioPsyKit>.

⁶*scikit-learn* (Version 1.6.1), available at <https://github.com/scikit-learn/scikit-learn>.

Chapter 6

Results

6.1 Statistical Analysis

In Figure 6.1, 6.2, 6.3, and 6.4 different pairs of features are plotted in such a way that the y-axis is shared across each row, and the x-axis across each column. An exception are the plots of the diagonal, which show a univariate distribution of the data of each column. In these plots, also the group with an indeterminate depression status is included.

In Table 6.1, 6.2, 6.3, and 6.4, the results of the statistical analysis for each of the feature domains are presented, where the ones with the indeterminate state are left out. As all of the features per domain were not normally distributed, the Mann-Whitney U test was applied to calculated all p-values. The p-values and effect sizes are interpreted according to Table 5.2.

On the diagonal of Figure 6.1, which represents different combinations of voice features, the distribution of healthy, depressed, and indeterminate for each chosen feature matches approximately, except for valence, where a slight distance between the distribution peaks can be seen. The pair-wise feature plots on the off-diagonal elements seem to have no general pattern in their distribution.

The results of the statistical analysis of voice features in Table 6.1 seem to confirm these observations. Valence has lowest value of $p = 0.029$ with a large effect size of 0.5, being the only feature in this domain with a significant difference between both groups. Arousal and dominance still both exhibit comparably low p-values as well (< 0.1) combined with medium effect sizes (around 0.4). With p-values in the range of 0.45 to 0.65, shimmer, jitter, and pitch have notably higher values compared to the three others.

In Figure 6.2, where distributions of different gait features are depicted, step length and gait velocity show a relatively normal distribution in the diagonal and tend to have separable group means. The samples of the indeterminate group are roughly in between these groups. There

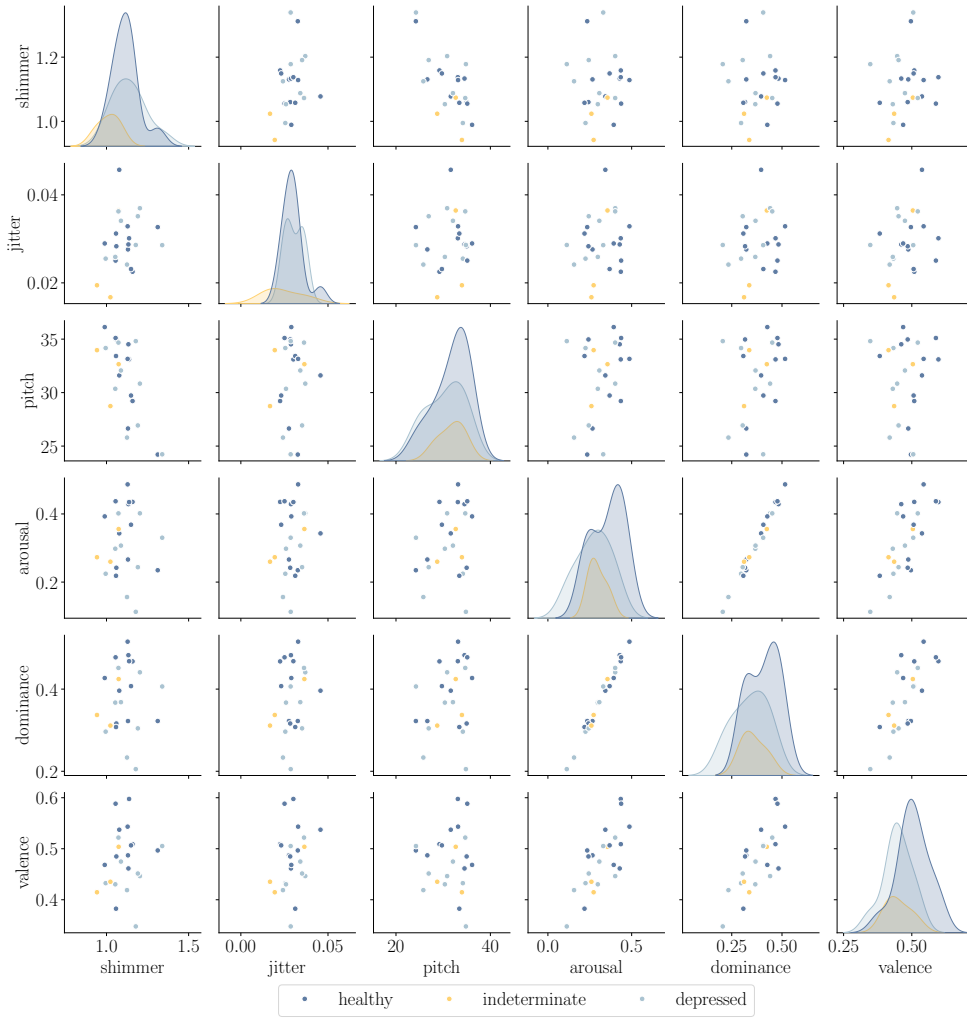


Figure 6.1: Distribution of the selected voice features in a pair plot.

Feature	Healthy	Depressed	U-statistic	p-value	Effect Size
Shimmer	1.12 ± 0.08	1.14 ± 0.10	47	0.644	-0.11
Jitter	0.03 ± 0.01	0.03 ± 0.01	46	0.594	-0.12
Pitch	31.82 ± 3.67	30.43 ± 3.97	65	0.456	0.17
Arousal	0.36 ± 0.09	0.28 ± 0.10	78	0.095	0.37
Dominance	0.41 ± 0.08	0.34 ± 0.09	81	0.060	0.42
Valence	0.50 ± 0.06	0.45 ± 0.05	86	0.025	0.50

Table 6.1: Results of statistical analysis showing mean $\pm$ standard deviation per group and test parameters of selected voice features.

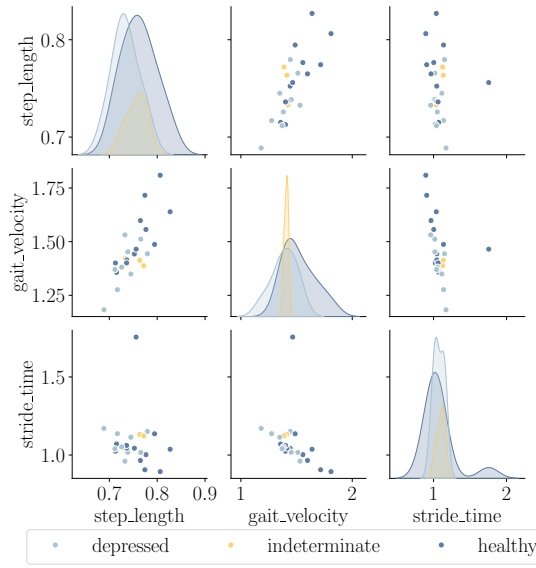


Figure 6.2: Distribution of the selected gait features in a pairplot.

Feature	Healthy	Depressed	U-statistic	p-value	Effect Size
Step length	0.76 ± 0.04	0.73 ± 0.03	78	0.095	0.37
Gait velocity	1.52 ± 0.14	1.39 ± 0.11	82	0.051	0.43
Stride time	1.08 ± 0.22	1.07 ± 0.07	41	0.374	-0.20

Table 6.2: Results of statistical analysis showing mean $\pm$ standard deviation per group and test parameters of selected gait features.

is a conspicuous outlier in stride time in the healthy group at approximately 1.8, leading to a non-normal distribution. All other samples – healthy, depressed, and indeterminate – have a stride time around 1.

The analysis, listed in Table 6.2, gives a similar picture. Gait velocity with a value of $p = 0.051$ exceeds almost the threshold towards being significantly different, while step length with $p = 0.095$ indicates a slight tendency towards this threshold. Both of those values come along with a medium effect sizes (around 0.4). Stride time exhibits a notably higher p-value than the others.

The distribution in the motion domain exhibit a similar picture both after the gait analysis shown in Figure 6.3 and during the interview shown in 6.4. The ratio of static periods with regard to moving periods and the mean duration of these static periods is higher in the depressed group, while the mean movement is greater in the healthy group. The entropy and absolute energy are

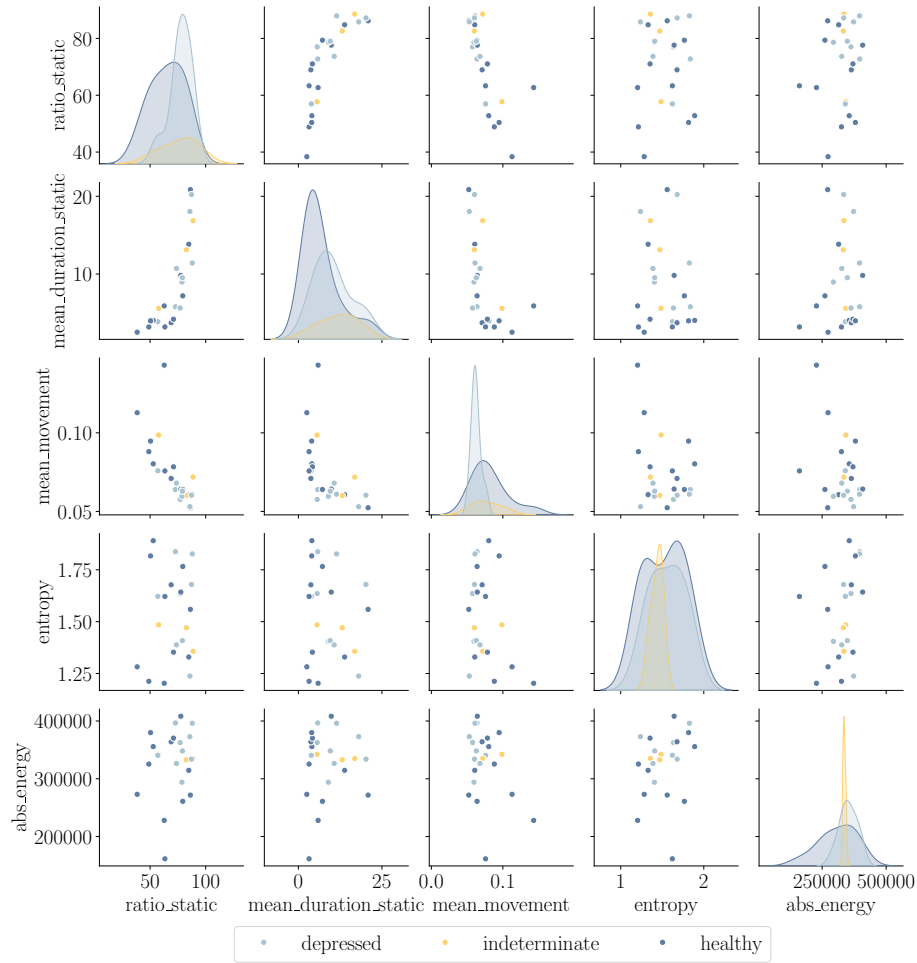


Figure 6.3: Distribution of the selected motion features, recorded after walking, in a pairplot.

Feature	Healthy	Depressed	U-statistic	p-value	Effect Size
Ratio static	65.37 ± 15.36	77.67 ± 9.58	27	0.060	-0.42
Mean duration static	6.87 ± 5.50	10.47 ± 5.54	30	0.095	-0.37
Mean movement	0.08 ± 0.02	0.06 ± 0.01	88	0.017	0.53
Entropy	1.53 ± 0.24	1.56 ± 0.21	46	0.594	-0.12
Absolute energy ($\times 10^6$)	0.31 ± 0.07	0.35 ± 0.03	35	0.189	-0.30

Table 6.3: Results of statistical analysis showing mean $\pm$ standard deviation per group and test parameters of selected motion features, recorded after walking.

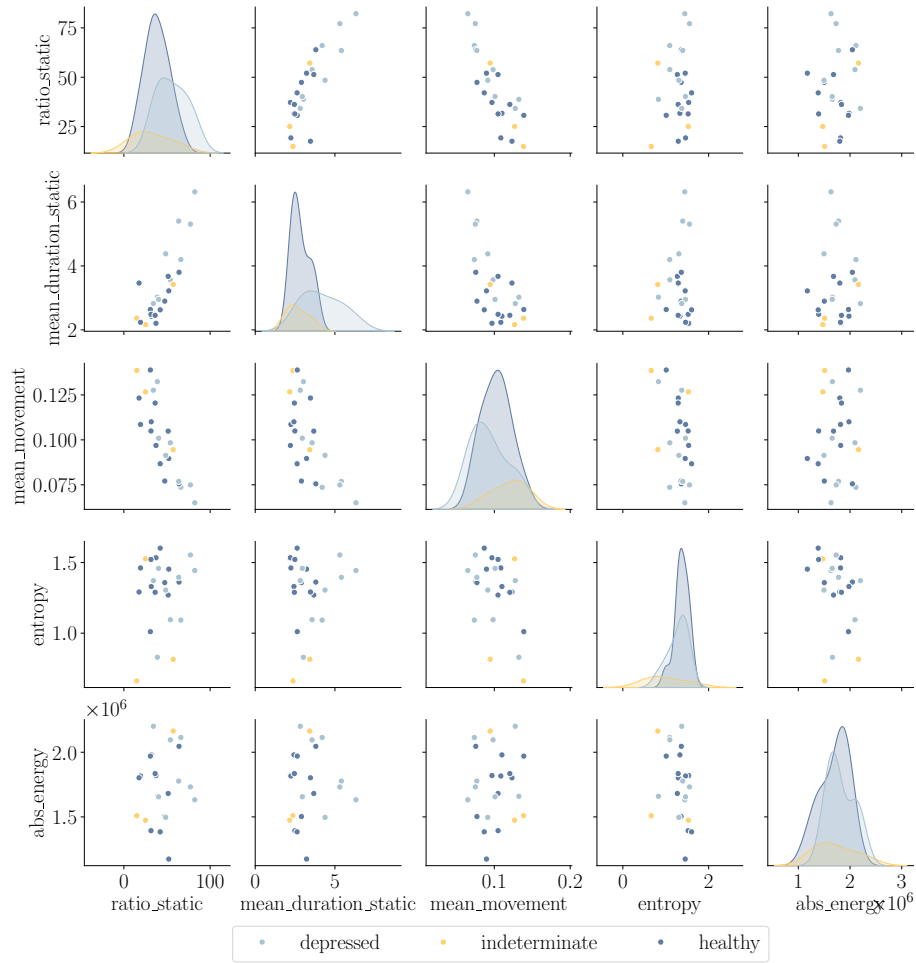


Figure 6.4: Distribution of the selected motion features, recorded during the interview, in a pairplot.

Feature	Healthy	Depressed	U-statistic	p-value	Effect Size
Ratio static	38.42 ± 13.68	56.06 ± 17.22	22	0.025	-0.50
Mean duration static	2.84 ± 0.56	4.22 ± 1.25	15	0.006	-0.60
Mean movement	0.10 ± 0.02	0.09 ± 0.02	71	0.241	0.26
Entropy	1.37 ± 0.16	1.28 ± 0.23	62	0.594	0.12
Absolute energy ($\times 10^6$)	1.70 ± 0.28	1.82 ± 0.25	47	0.644	-0.11

Table 6.4: Results of statistical analysis showing mean $\pm$ standard deviation per group and test parameters of selected motion features, recorded during the interview.

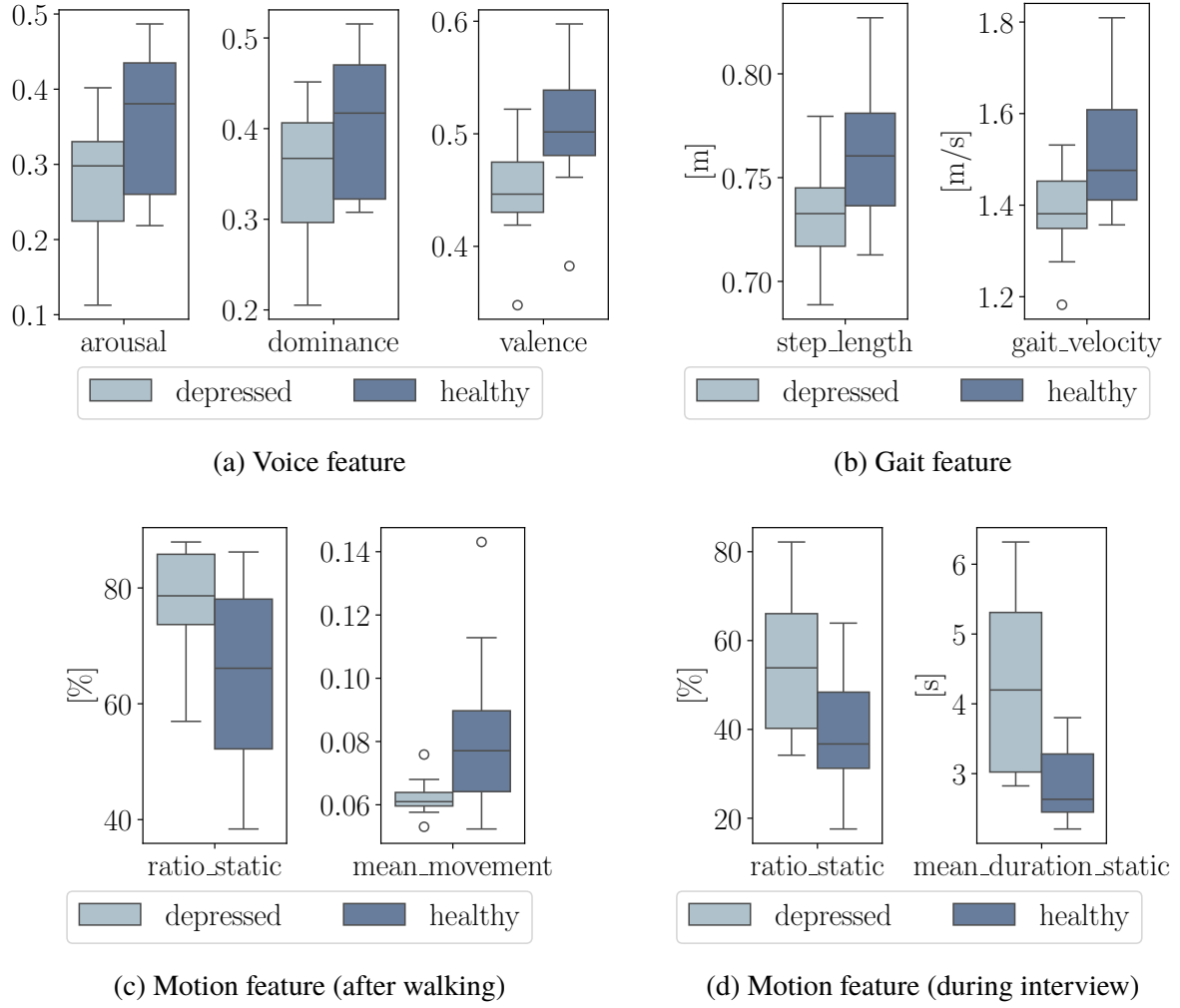


Figure 6.5: Selection of features with low p-values per domain showing the locality of their values across both groups with box plots.

both more or less equal across groups during both recording settings.

These observations also largely coincide with the results of the statistical analysis. In the recording after walking, which is shown in Table 6.3, the mean movement exhibits the lowest p-value of 0.017 along with a large effect size (> 0.5). The ratio and mean duration of static periods both are also comparably low (around 0.060 and 0.095), but only exhibit medium effect sizes. During the interview, the p-values of both of these are the lowest within this domain alongside with large effect sizes, while mean movement, entropy, and absolute energy exhibit notably higher p-values accompanied by small effect sizes.

The distribution of those features, that have a low p-value compared to other features in their domain, are additionally visualized in the plots of Figure 6.5.

6.2 Classification Experiments

Table 6.5 summarizes the classification results across different feature domains and classifiers. Each domain corresponds to a specific combination of sensor modalities: the *voice* and *gait* domains include only vocal and gait-related features, respectively. The *interview* domain combines both voice and head movement features captured during the interview session, while the *IMU* domain aggregates motion features from the gait task, the subsequent motion task, and the movements captured during the interview. Lastly, the *combined* domain incorporates all available features across modalities.

Balanced accuracy served as the primary evaluation metric, while precision, recall, and F1 score were reported as additional performance indicators.

In the voice domain, classification performances were in a range of $\pm 20\%$. The best result was achieved by the kNN classifier, with a balanced accuracy of 0.62 ± 0.23 . Other models performed near or below chance level, such as LR and SVM. A similar trend was observed in the gait domain, where LR yielded the highest balanced accuracy of 0.62 ± 0.11 , followed by kNN (0.57 ± 0.25). Other classifiers, particularly SVM, performed notably worse.

Performance improved substantially in the multi-domain settings. The interview-based features brought performance improvements across all classifiers, with RF achieved the highest balanced accuracy (0.80 ± 0.29), while kNN yielded the highest precision and F1 score and matched the RF's recall. Similarly, in the IMU domain, RF reached a balanced accuracy of 0.80 ± 0.11 , also showing identical results for precision, recall, and F1 score. LR, kNN, and XGBoost closely followed, whereas SVM demonstrated a slightly weaker performance.

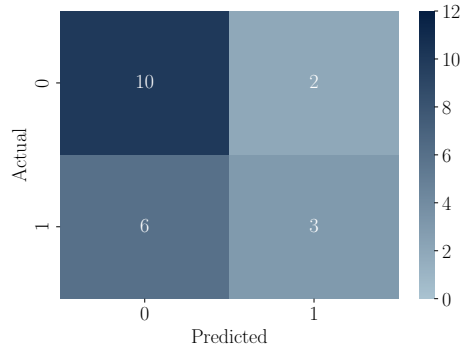
The highest overall performance was obtained using the combined feature set. Again, the RF classifier achieved the best results across all metrics, with a balanced accuracy of 0.85 ± 0.13 . LR achieved results comparable to its performance in the IMU domain, while SVM performance deteriorated further.

Figure 6.6 presents the confusion matrix and the SHAP summary plots for one of the best-performing classifiers in each domain. In the SHAP plots, features are ranked by their overall importance on the y-axis. The x-axis represents the SHAP values, indicating the magnitude and direction of each feature's contribution to the model's output – positive values suggest an increase in the prediction, while negative values indicate a decrease. The color gradient reflects whether high or low feature values in red and blue respectively are associated with the increase or decrease of the prediction.

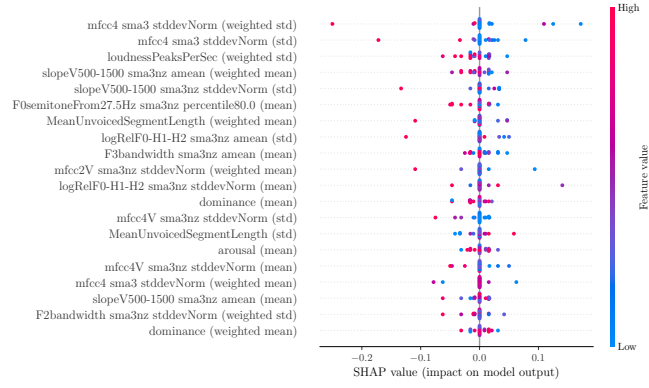
In the voice domain (Figure 6.6a), the kNN classifier showed a bias toward classifying participants as non-depressed by misclassifying 6 out of 16 in this case. SHAP analysis (Figure 6.6b)

Domain	Classifier	Balanced Accuracy	Precision	Recall	F1 Score
Voice	LR	0.43 ± 0.13	0.31 ± 0.27	0.43 ± 0.42	0.35 ± 0.32
	kNN	0.62 ± 0.23	0.57 ± 0.35	0.62 ± 0.37	0.57 ± 0.33
	RF	0.57 ± 0.21	0.49 ± 0.33	0.57 ± 0.35	0.49 ± 0.28
	SVM	0.43 ± 0.10	0.29 ± 0.26	0.43 ± 0.42	0.34 ± 0.31
	XGBoost	0.58 ± 0.21	0.53 ± 0.28	0.58 ± 0.31	0.55 ± 0.28
Gait	LR	0.62 ± 0.11	0.57 ± 0.33	0.62 ± 0.37	0.56 ± 0.30
	kNN	0.57 ± 0.25	0.50 ± 0.36	0.57 ± 0.39	0.49 ± 0.32
	RF	0.47 ± 0.25	0.41 ± 0.36	0.47 ± 0.42	0.42 ± 0.35
	SVM	0.43 ± 0.24	0.38 ± 0.35	0.43 ± 0.42	0.38 ± 0.34
	XGBoost	0.50 ± 0.19	0.42 ± 0.27	0.50 ± 0.32	0.43 ± 0.25
Interview	LR	0.67 ± 0.19	0.45 ± 0.33	0.55 ± 0.41	0.42 ± 0.47
	kNN	0.75 ± 0.16	0.76 ± 0.30	0.75 ± 0.34	0.72 ± 0.26
	RF	0.80 ± 0.29	0.72 ± 0.32	0.75 ± 0.34	0.71 ± 0.31
	SVM	0.75 ± 0.16	0.76 ± 0.31	0.75 ± 0.34	0.71 ± 0.27
	XGBoost	0.68 ± 0.24	0.62 ± 0.37	0.68 ± 0.38	0.62 ± 0.34
IMU	LR	0.80 ± 0.19	0.77 ± 0.33	0.80 ± 0.33	0.75 ± 0.29
	kNN	0.80 ± 0.19	0.77 ± 0.33	0.80 ± 0.33	0.75 ± 0.29
	RF	0.80 ± 0.11	0.83 ± 0.21	0.80 ± 0.26	0.76 ± 0.15
	SVM	0.78 ± 0.19	0.77 ± 0.33	0.78 ± 0.35	0.73 ± 0.30
	XGBoost	0.80 ± 0.19	0.77 ± 0.33	0.80 ± 0.33	0.75 ± 0.29
Combined	LR	0.80 ± 0.19	0.77 ± 0.33	0.80 ± 0.33	0.75 ± 0.29
	kNN	0.80 ± 0.19	0.77 ± 0.33	0.80 ± 0.33	0.75 ± 0.29
	RF	0.85 ± 0.13	0.87 ± 0.21	0.85 ± 0.24	0.81 ± 0.17
	SVM	0.65 ± 0.20	0.56 ± 0.35	0.65 ± 0.39	0.57 ± 0.33
	XGBoost	0.77 ± 0.16	0.72 ± 0.33	0.77 ± 0.33	0.70 ± 0.27

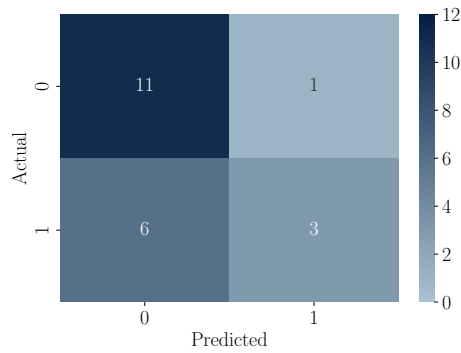
Table 6.5: Performance (mean $\pm$ standard deviation) of different classifiers across the different feature domains. The best results per domain are highlighted in bold.



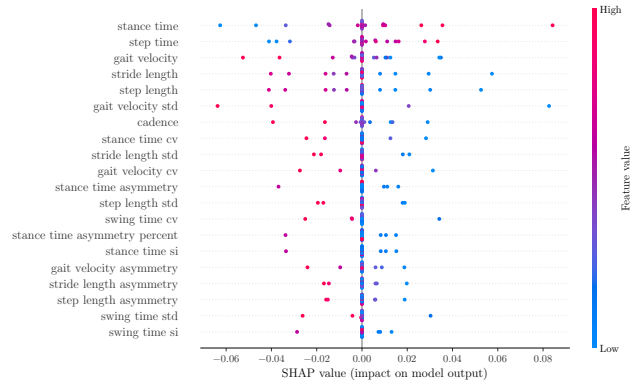
(a) Confusion matrix (voice)



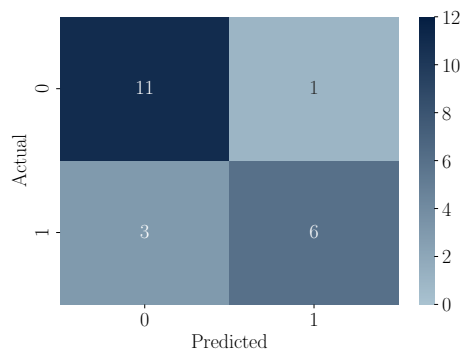
(b) SHAP values (voice)



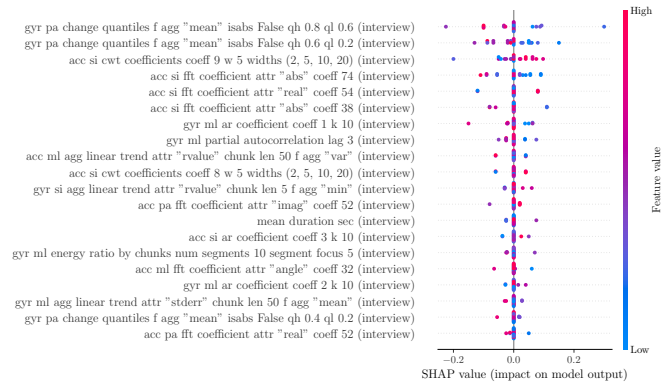
(c) Confusion matrix (gait)



(d) SHAP values (gait)

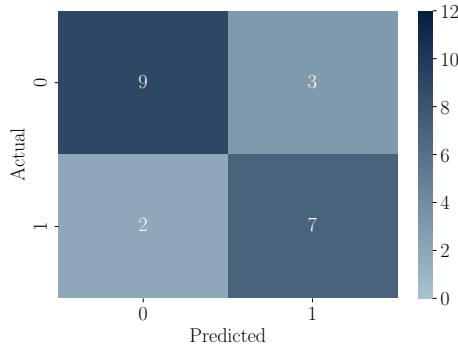


(e) Confusion matrix (interview)

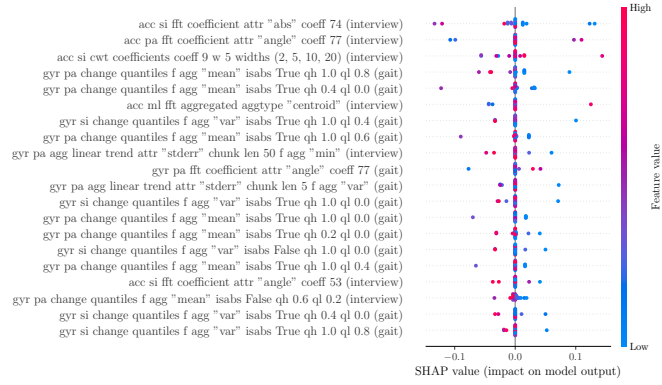


(f) SHAP values (interview)

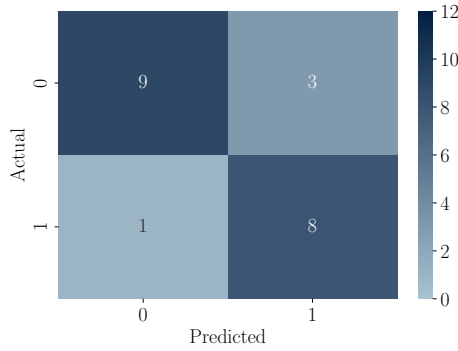
Figure 6.6: Analysis of the best performing classification pipelines of the different domains.



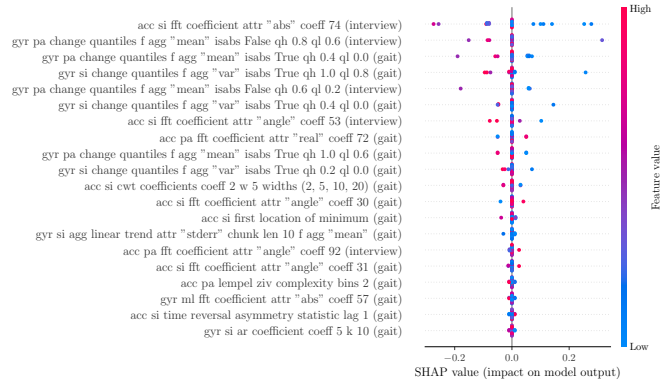
(g) Confusion matrix (IMU)



(h) SHAP values (IMU)



(i) Confusion matrix (combined)



(j) SHAP values (combined)

Figure 6.6: Analysis of the best performing classification pipelines of the different domains.

indicated that features related to MFCCs (e. g. *mfcc4 sma3*), spectral (e. g. *logRelF0-H1-H2*), and temporal characteristics (e. g. *loudnessPeaksPerSec*), as well as emotional dimensions *arousal* and *dominance* were among the most influential.

The analysis of the gait classification pipeline shows a similar result in the confusion matrix (Figure 6.6c). Key features identified in the SHAP analysis (Figure 6.6d) included *stance time* and *gait velocity* as well as *step length* measures.

Figure 6.6e demonstrates improved detection of depressed individuals using all features from the interview session. RF correctly classified 6 out of 9 depressed cases, and SHAP analysis (Figure 6.6f) showed the top features were dominated by motion descriptors, including generic and expert features (e.g. *mean duration sec*), with no vocal features appearing among the top 20.

A similar improvement can be seen in Figure 6.6g. SHAP analysis (Figure 6.6h) revealed a mixture of both motion modalities including head movements during the interview and after the walking part (marked with (*gait*)).

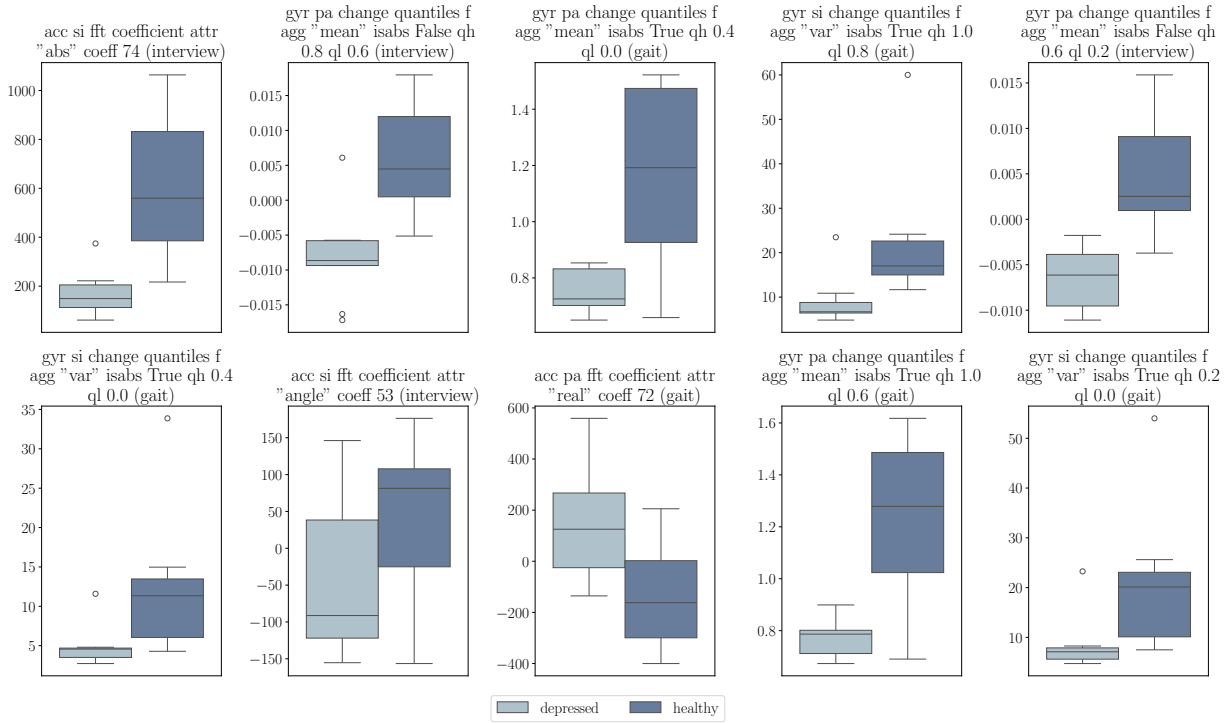


Figure 6.7: Box plots of the ten features ranked highest by the SHAP analysis of the best-performing classifier (see also Figure 6.6j).

Finally, the combined domain (Figure 6.6i) provided the most balanced results, accurately identifying both depressed and non-depressed participants. The corresponding SHAP analysis (Figure 6.6j) showed that all top 20 most influential features originated from motion data, which is coherent to what is observed in the interview and IMU domain (see Figure 6.6f and 6.6h).

To further assess whether the top-ranked features identified through SHAP analysis genuinely possess discriminative power, the ten highest-ranking features from the best-performing classifier were extracted and their distributions across the two groups (depressed and healthy) were visualized using box plots, as shown in Figure 6.7. This serves as a sanity check to validate the model's behavior by inspecting whether the features that most strongly influenced classification decisions also exhibit meaningful group-level differences. The results indicate that these features effectively separate the two groups, supporting their relevance in distinguishing depressive symptomatology.

Chapter 7

Discussion

This chapter interprets and contextualizes the findings of the present study, which explored the feasibility of using ear-worn devices to detect depressive symptoms through behavioral markers in gait, motion, and voice. By applying both statistical analyses and machine learning techniques to a newly collected dataset, the study aimed to identify meaningful differences between healthy and depressed individuals and evaluate the potential of multi-modal sensor data for mental health monitoring. The following sections discuss the significance of these findings, compare them with existing literature, and outline the implications and limitations of the proposed approach.

Due to the exploratory nature of this study, a statistical analysis was performed to gain insights into the newly recorded dataset. The analysis revealed several behavioral differences between healthy and depressed individuals, particularly within gait and motion features. In the voice domain, the emotional dimensions – especially valence, which had the lowest p-value, followed by arousal and dominance – emerged as potential indicators of depressive states. This suggests that emotional tone in speech may serve as a useful marker for affective well-being. In contrast, the low level descriptors (LLD) from the GeMAPS feature set exhibited relatively high p-values, indicating limited discriminative power in this context. A possible explanation for that may lie in the small sample size and the lack of gender stratification. Unlike other comparable studies [Fra00; Jia18], this analysis did not differentiate between male and female participants, which could have influenced features sensitive to gender differences, such as F_0 .

In the gait domain, gait velocity produced the lowest p-value, confirming earlier findings that associate reduced walking speed with depressive symptoms [Lem00]. Although not all previously reported results could be replicated – potentially due to higher error rates in gait event detection from ear-worn IMUs compared to other systems [Sei24] – a general trend consistent with earlier research was observable. Participants with depression also showed reduced overall activity

and more static motion, particularly during interviews, aligning with previous observations of psychomotor slowing in depressive states [OBr17; Gha17].

The classification experiments further contextualize these results. When evaluating single-domain models, moderate classification performance was achieved. In the voice domain, a kNN classifier yielded a balanced accuracy of 0.62 ± 0.23 , while LR in the gait domain reached 0.62 ± 0.11 . These results are in line with similar approaches in the literature. For instance, Shin et al. [Shi21] achieved approximately 60 % classification accuracy using voice features extracted from semi-structured interviews, and Zhao et al. [Zha19] obtained a 64 % accuracy using camera-recorded gait data to distinguish between depressed and healthy individuals. However, these outcomes fall short of the higher accuracies reported in studies with larger and more homogeneous datasets. [Par17], for example, used the DAIC dataset and custom-engineered voice features to train ML models that achieved over 90 %. A key factor likely contributing to this difference is the larger sample size ($n = 85$ in [Par17] versus $n = 21$ in this study), along with more refined pre-processing and feature selection pipelines. Likewise, studies employing IMU-based gait analysis have reported accuracies exceeding 85 % [Jun22; Boo23], supported by significantly larger number of participants. These studies again exhibited notably higher sample sizes, for example of $n = 133$ [Boo23]. They also often focused on more homogeneous groups, such as university students [Boo23] or community-dwelling elderly [Jun22], whereas the current dataset included a broader age range (21 to 72 years), likely increasing within-group variability.

Importantly, combining features across domains led to substantial improvements in classification performance. Multi-modal models integrating voice and motion (*interview* domain) or gait and motion (*IMU* domain), achieved accuracies of up to 80 %. The highest performance with a balanced accuracy of 0.85 ± 0.13 was obtained when all available features were used. This finding supports previous research showing that multi-modal fusion improves predictive performance [Dai23; Li25]. In all cases, RF classifiers outperformed others, likely due to their capacity to model complex inter-domain relationships. Although a new set of hyperparameters was selected in each outer fold of the nested CV, RF classifiers consistently showed promise for multi-modal depression classification.

The application of ML models was particularly valuable given that traditional statistical approaches based solely on literature-derived markers failed to reliably differentiate between groups. SHAP analysis of the best-performing classifier highlighted specific features with high discriminative potential. However, due to the small number of participants relative to the number of extracted features – particularly in the motion domain – these rankings must be interpreted with caution, as they may not generalize across different datasets or populations. Nevertheless,

motion-related features emerged consistently as top contributors across domains, reinforcing their relevance and potential in depression detection.

From a practical standpoint, the use of ear-worn devices such as HAs and earbuds offers a promising avenue for real-world applications, especially because hearing loss is a potential risk factor for depression [Sha21]. These devices integrate naturally into users' daily routines and can unobtrusively capture multi-modal data, including voice, motion, and contextual activity. Despite this potential, the current approach is not yet suitable for real-time applications due to the computational demands of pre-processing steps, such as running pretrained DL models and storing longitudinal data. However, streaming data to smartphones for real-time processing may offer a viable approach to preserve the method's applicability in everyday settings.

Although the study was designed to reflect naturalistic behavior, it was recorded within a laboratory environment and time-constraint setting, both of which may limit the generalizability to real-life use cases. Additionally, the small sample size restricts the statistical power of the findings and precludes a fully balanced study design. For instance, the average age in the depressed group was lower than in the healthy group, introducing potential confounds. However, findings such as higher gait velocity among the older healthy participants – a reversal of the typical age-related decline in speed [Hol11] – suggest that some behavioral differences are likely attributable to mental health status rather than demographic variation alone.

In summary, this study demonstrates that HA-integrated sensors can provide a feasible and non-intrusive method for identifying individuals with depressive symptoms, particularly when capturing naturalistic behaviors during tasks such as interviews or walking. The results highlight the diagnostic value of psychomotor features and advantage of multi-modal data fusion. While limitations remain, especially regarding sample size and real-world applicability, the findings lay a promising foundation for future work in wearable-based mental health monitoring and early intervention.

Chapter 8

Conclusion and Outlook

This thesis investigated the feasibility of detecting depressive symptoms using HA-integrated microphones and IMUs by analyzing voice, gait, and motion features. To support this investigation, a novel dataset was recorded, comprising interviews and sensor data from 25 participants. Based on their PHQ-8 scores, three participants were excluded due to discrepancies between their initial screening and on-site assessment. Twelve individuals were categorized as healthy controls ($\text{PHQ-8} < 10$) and nine as depressed ($\text{PHQ-8} \geq 10$), forming the final dataset of 21 participants for analysis.

A range of behavioral characteristics and sensor modalities were examined, including acoustic voice features, gait parameters, and both generic and expert-informed motion features derived from IMU signals. Both statistical analysis and ML experiments demonstrated that these features provided reliable indicators for distinguishing between the two groups. Notably, the multi-modal fusion of sensor data significantly improved classification performance. A random forest classifier trained on all available features achieved a balanced accuracy of $85 \pm 13\%$, with IMU-derived motion data contributing most prominently across domains.

These findings highlight the practical viability of utilizing a single wearable device, a HA inherently integrating a microphone and IMU sensors, to capture multiple behavioral modalities relevant to depression detection. The proposed approach demonstrates competitive performance relative to state-of-the-art systems while offering a more passive and unobtrusive solution. Although further validation in larger-scale studies is required, this work represents a significant step toward the development of wearable systems capable of detecting depression in real-world, daily-life settings.

Future research should address several limitations and explore directions to enhance the robustness and applicability of HA-based depression detection systems. Increasing the number and diversity of participants would improve generalizability and help reduce the high variance

observed in the current ML model results. Furthermore, while this study employed a largely data-driven approach for feature extraction in the motion domain, a more comprehensive and targeted integration of domain-specific knowledge could help identify motion patterns that are not only predictive but also clinically meaningful, enhancing both interpretability and relevance. In parallel, future studies should collect data during everyday routines to move beyond laboratory settings. Monitoring behavior in natural contexts brings the system closer to real-world mental health applications. Alternative modeling approaches – such as DL architectures, advanced sensor fusion techniques, or regression-based frameworks – could be explored to improve performance or enable symptom severity estimation. Finally, the proposed framework could be adapted to monitor other mental health disorders, such as anxiety or bipolar disorder, expanding the potential scope and impact of wearable sensing in mental health.

In summary, this work establishes a compelling foundation for the use of everyday wearable technology to passively monitor mental health. Through a combination of psychomotor indicators, ML techniques, and multi-modal sensor fusion, HA-integrated systems could support scalable, early-stage depression detection and long-term tracking of mental well-being in real-world settings.

List of Figures

2.1	Overview of the vocal tract, adapted from from [Do 25].	6
2.2	Functional divisions of the gait cycle according to [Per10], Figure adapted from [Stö15].	9
2.3	Illustration of measures within a gait cycle: stride and step length, step width, and foot angle. Figure adapted from [Sto98].	10
2.4	Gait analysis pipeline, adapted from [Küd24].	11
2.5	Training and testing process, adapted from [Uça20].	12
2.6	Sampling principle of a random forest. Adapted from [Sir20].	15
2.7	A separating linear hyperplane with margin γ for data from two classes (squares, circles), adapted from [Mam09].	16
2.8	The Transformer model architecture, adapted from [Vas17].	18
4.1	Experimental setup of the interview during the study (situation was recreated). . .	30
5.1	Process of getting speech segments of the participant.	32
5.2	Sensor position during the interview.	34
6.1	Distribution of the selected voice features in a pair plot.	40
6.2	Distribution of the selected gait features in a pairplot.	41
6.3	Distribution of the selected motion features, recorded after walking, in a pairplot. .	42
6.4	Distribution of the selected motion features, recorded during the interview, in a pairplot.	43
6.5	Selection of features with low p-values per domain showing the locality of their values across both groups with box plots.	44
6.6	Analysis of the best performing classification pipelines of the different domains. .	47
6.6	Analysis of the best performing classification pipelines of the different domains. .	48

6.7	Box plots of the ten features ranked highest by the SHAP analysis of the best-performing classifier (see also Figure 6.6j).	49
-----	---	----

List of Tables

4.1	Inclusion and exclusion criteria for study participation.	28
4.2	Sociodemographic data of the participants, further subdivided by gender. Age, height and weight is specified with mean $\pm$ standard deviation. Difference to group “All” due to the four exclusions.	29
5.1	Microphone channels and their corresponding gain corrections.	31
5.2	Overview of statistical tests and effect size measures used in the analysis.	36
6.1	Results of statistical analysis showing mean $\pm$ standard deviation per group and test parameters of selected voice features.	40
6.2	Results of statistical analysis showing mean $\pm$ standard deviation per group and test parameters of selected gait features.	41
6.3	Results of statistical analysis showing mean $\pm$ standard deviation per group and test parameters of selected motion features, recorded after walking.	42
6.4	Results of statistical analysis showing mean $\pm$ standard deviation per group and test parameters of selected motion features, recorded during the interview.	43
6.5	Performance (mean $\pm$ standard deviation) of different classifiers across the different feature domains. The best results per domain are highlighted in bold.	46
C.1	Grid-search parameter values for feature selection and classification algorithms. .	108

Bibliography

- [Alg12] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Michael Breakspear, Gordon Parker, et al. “From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech”. In: *FLAIRS*. 2012.
- [Alg13] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Tom Gedeon, Michael Breakspear, and Gordon Parker. “A Comparative Study of Different Classifiers for Detecting Depression from Spontaneous Speech”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 8022–8026. doi: 10.1109/ICASSP.2013.6639227.
- [Ath19] Chanavit Athavipach, Setha Pan-ngum, and Pasin Israsena. “A Wearable In-Ear EEG Device for Emotion Monitoring”. In: *Sensors* 19.18 (2019), p. 4014. doi: 10.3390/s19184014.
- [Bae20] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 12449–12460.
- [Bar23] Hika Barki and Wan-Young Chung. “Mental Stress Detection Using a Wearable In-Ear Plethysmography”. In: *Biosensors* 13.3 (2023), p. 397. doi: 10.3390/bios13030397.
- [Bec96] A. T. Beck, R. A. Steer, R. Ball, and W. Ranieri. “Comparison of Beck Depression Inventories -IA and -II in Psychiatric Outpatients”. In: *Journal of Personality Assessment* 67.3 (1996), pp. 588–597. doi: 10.1207/s15327752jpa6703_13.
- [Ben01] Ralf Bender and Stefan Lange. “Adjusting for Multiple Testing—When and How?” In: *Journal of Clinical Epidemiology* 54.4 (2001), pp. 343–349. doi: 10.1016/S0895-4356(00)00314-0.

- [Ber12] Matthias Berking and Winfried Rief. “Affektive Störungen und Suizidalität”. In: *Klinische Psychologie und Psychotherapie für Bachelor*. Ed. by Matthias Berking and Winfried Rief. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 29–67. ISBN: 978-3-642-16973-1.
- [Bho21] Achintya K. Bhowmik, David A. Fabry, Phillip Armour, Hal Berghel, Robert N. Charette, and John L. King. “Hear, Now, and in the Future: Transforming Hearing Aids Into Multipurpose Devices”. In: *Computer* 54.11 (2021), pp. 108–120. doi: 10.1109/MC.2021.3105151.
- [Bi21] Jiarui Bi, Zengliang Zhu, and Qinglong Meng. “Transformer in Computer Vision”. In: *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*. 2021, pp. 178–188. doi: 10.1109/CEI52496.2021.9574462.
- [Bia16] Gérard Biau and Erwan Scornet. “A Random Forest Guided Tour”. In: *TEST* 25.2 (2016), pp. 197–227. doi: 10.1007/s11749-016-0481-7.
- [Bis06] Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. New York: Springer, 2006. ISBN: 978-0-387-31073-2.
- [Bon21] Erika Bondareva, Elín Rós Hauksdóttir, and Cecilia Mascolo. “Earables for Detection of Bruxism: a Feasibility Study”. 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*. Virtual USA: ACM, 2021, pp. 146–151. ISBN: 978-1-4503-8461-2. doi: 10.1145/3460418.3479327.
- [Boo23] Ali Boolani, Allison H. Gruber, Ahmed Ali Torad, and Andreas Stamatis. “Identifying Current Feelings of Mild and Moderate to High Depression in Young, Healthy Individuals Using Gait and Balance: An Exploratory Study”. In: *Sensors* 23.14 (2023), p. 6624. doi: 10.3390/s23146624.
- [Bou14] Loubna Bouarfa, Louis Atallah, Richard Mark Kwasnicki, Claire Pettitt, Gary Frost, and Guang-Zhong Yang. “Predicting Free-Living Energy Expenditure Using a Miniaturized Ear-Worn Sensor: An Evaluation Against Doubly Labeled Water”. In: *IEEE Transactions on Biomedical Engineering* 61.2 (2014), pp. 566–575. doi: 10.1109/TBME.2013.2284069.
- [Bre01] Leo Breiman. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32. doi: 10.1023/A:1010933404324.

- [Bre23] Hervé Bredin. “pyannote.audio 2.1 Speaker Diarization Pipeline: Principle, Benchmark, and Recipe”. In: *INTERSPEECH 2023*. ISCA, 2023, pp. 1983–1987. doi: 10.21437/Interspeech.2023-105.
- [Buk13] Jens Drachmann Bukh, Camilla Bock, Maj Vinberg, and Lars Vedel Kessing. “The Effect of Prolonged Duration of Untreated Depression on Antidepressant Treatment Outcome”. In: *Journal of Affective Disorders* 145.1 (2013), pp. 42–48. doi: 10.1016/j.jad.2012.07.008.
- [Cha02] Henry G. Chambers and David H. Sutherland. “A Practical Guide to Gait Analysis”. In: *JAAOS - Journal of the American Academy of Orthopaedic Surgeons* 10.3 (2002), p. 222.
- [Che16] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. doi: 10.1145/2939672.2939785.
- [Chl19] Karol Chlasta, Krzysztof Wołk, and Izabela Krejtz. “Automated Speech-Based Screening of Depression Using Deep Convolutional Neural Networks”. In: *Procedia Computer Science*. CENTERIS 2019 - International Conference on ENTERprise Information Systems / ProjMAN 2019 - International Conference on Project MANagement / HCist 2019 - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2019 164 (2019), pp. 618–628. doi: 10.1016/j.procs.2019.12.228.
- [Chr18] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. “Time Series Feature Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package)”. In: *Neurocomputing* 307 (2018), pp. 72–77. doi: 10.1016/j.neucom.2018.03.067.
- [Coh13] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. New York: Routledge, 2013. ISBN: 978-0-203-77158-7.
- [Cou97] John L. Coulehan, Herbert C. Schulberg, Marian R. Block, Michael J. Madonia, and Eric Rodriguez. “Treating Depressed Primary Care Patients Improves Their Physical, Mental, and Social Functioning”. In: *Archives of Internal Medicine* 157.10 (1997), pp. 1113–1120. doi: 10.1001/archinte.1997.00440310079008.

- [Cum15] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F. Quatieri. “A Review of Depression and Suicide Risk Assessment Using Speech Analysis”. In: *Speech Communication* 71 (2015), pp. 10–49. doi: 10.1016/j.specom.2015.03.004.
- [Cur56] Edward E. Cureton. “Rank-Biserial Correlation”. In: *Psychometrika* 21.3 (1956), pp. 287–290. doi: 10.1007/BF02289138.
- [Dai23] Ziqian Dai, Qiuping Li, Yichen Shang, and Xin’an Wang. “Depression Detection Based on Facial Expression, Audio and Gait”. In: *2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. Chongqing, China: IEEE, 2023, pp. 1568–1573. ISBN: 978-1-6654-6004-0. doi: 10.1109/ITNEC56291.2023.10082163.
- [Den01] Peter B. Denes and Elliot N. Pinson. *The Speech Chain: The Physics and Biology of Spoken Language*. 8. print. New York, N.Y: Freeman, 2001. ISBN: 978-0-7167-2256-4.
- [Dia20] Yanan Diao, Yu Ma, Dongyan Xu, Wei Chen, and Yuanyuan Wang. “A Novel Gait Parameter Estimation Method for Healthy Adults and Postoperative Patients With an Ear-Worn Sensor”. In: *Physiological Measurement* 41.5 (2020), 05NT01. doi: 10.1088/1361-6579/ab87b5.
- [Día20] Steven Díaz, Jeannie B. Stephenson, and Miguel A. Labrador. “Use of Wearable Sensor Technology in Gait, Balance, and Range of Motion Analysis”. In: *Applied Sciences* 10.1 (2020), p. 234. doi: 10.3390/app10010234.
- [Do 25] Eriberto Do Nascimento. *Organic Production of Speech Sounds in Phonetics*. 2025. URL: <https://englishphonetics.net/english-phonetics-academy/production-speech-sounds-phonetics.html> (visited on 03/10/2025).
- [Dom12] Pedro Domingos. “A Few Useful Things to Know About Machine Learning”. In: *Communications of the ACM* 55.10 (2012), pp. 78–87. doi: 10.1145/2347736.2347755.
- [Esp21] Caroline Wanderley Espinola, Juliana Carneiro Gomes, Jessiane Mônica Silva Pereira, and Wellington Pinheiro dos Santos. “Detection of Major Depressive Disorder Using Vocal Acoustic Analysis and Machine Learning—An Exploratory Study”. In: *Research on Biomedical Engineering* 37.1 (2021), pp. 53–64. doi: 10.1007/s42600-020-00100-9.

- [Eyb10] Florian Eyben, Martin Wöllmer, and Björn Schuller. “Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor”. In: *Proceedings of the 18th ACM international conference on Multimedia*. MM ’10. New York, NY, USA: Association for Computing Machinery, 2010, pp. 1459–1462. ISBN: 978-1-60558-933-6. DOI: 10.1145/1873951.1874246.
- [Eyb16] Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing”. In: *IEEE Transactions on Affective Computing* 7.2 (2016), pp. 190–202. DOI: 10.1109/TAFFC.2015.2457417.
- [Fel20] Ron Feldman, Shaul Schreiber, Chaim Pick, and Ella Been. “Gait, Balance and Posture in Major Mental Illnesses: Depression, Anxiety and Schizophrenia”. In: (2020).
- [Fer94] F. Ferraris, I. Gorini, U. Grimaldi, and M. Parvis. “Calibration of three-axial rate gyros without angular velocity standards”. In: *Sensors and Actuators A: Physical*. Proceedings of Eurosensors VIII 42.1 (1994), pp. 446–449. DOI: 10.1016/0924-4247(94)80031-6.
- [Fix51] E. Fix and J.L. Hodges. *Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties*. USAF School of Aviation Medicine, 1951.
- [Fra00] D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes. “Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk”. In: *IEEE Transactions on Biomedical Engineering* 47.7 (2000), pp. 829–837. DOI: 10.1109/10.846676.
- [Fri01] Jerome H. Friedman. “Greedy Function Approximation: A Gradient Boosting Machine”. In: *The Annals of Statistics* 29.5 (2001), pp. 1189–1232.
- [Fri14] Eiko I. Fried and Randolph M. Nesse. “The Impact of Individual Depressive Symptoms on Impairment of Psychosocial Functioning”. In: *PLOS ONE* 9.2 (2014), e90311. DOI: 10.1371/journal.pone.0090311.
- [Ger23] Maurice Gerczuk, Andreas Triantafyllopoulos, Shahin Amiriparian, Alexander Kathan, Jonathan Bauer, Matthias Berking, and Björn W. Schuller. “Zero-Shot Personalization of Speech Foundation Models for Depressed Mood Monitoring”. In: *Patterns* 4.11 (2023), p. 100873. DOI: 10.1016/j.patter.2023.100873.

- [Gha17] Asma Ghandeharioun, Szymon Fedor, Lisa Sangermano, Dawn Ionescu, Jonathan Alpert, Chelsea Dale, David Sontag, and Rosalind Picard. “Objective Assessment of Depressive Symptoms with Machine Learning and Wearable Sensors Data”. In: *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. ISSN: 2156-8111. 2017, pp. 325–332. DOI: 10.1109/ACII.2017.8273620.
- [Gra14] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. “The Distress Analysis Interview Corpus of Human and Computer Interviews”. In: *LREC*. Vol. 14. Reykjavik, 2014, pp. 3123–3128.
- [Gru14] Agnes Gruenerbl, Venet Osmani, Gernot Bahle, Jose C. Carrasco, Stefan Oehler, Oscar Mayora, Christian Haring, and Paul Lukowicz. “Using Smart Phone Mobility Traces for the Diagnosis of Depressive and Manic Episodes in Bipolar Patients”. In: *Proceedings of the 5th Augmented Human International Conference*. Kobe Japan: ACM, 2014, pp. 1–8. ISBN: 978-1-4503-2761-9. DOI: 10.1145/2582051.2582089.
- [Guj19] Trupti Gujarathi and Kalyani Bhole. “Gait Analysis Using IMU Sensor”. In: *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 2019, pp. 1–5. DOI: 10.1109/ICCCNT45670.2019.8944545.
- [Ham80] M. Hamilton. “Rating Depressive Patients”. In: *The Journal of Clinical Psychiatry* 41.12 Pt 2 (1980), pp. 21–24.
- [Han24] Zhuojin Han, Yuanyuan Shang, Zhuhong Shao, Jingyi Liu, Guodong Guo, Tie Liu, Hui Ding, and Qiang Hu. “Spatial–Temporal Feature Network for Speech-Based Depression Recognition”. In: *IEEE Transactions on Cognitive and Developmental Systems* 16.1 (2024), pp. 308–318. DOI: 10.1109/TCDS.2023.3273614.
- [He18] Lang He and Cui Cao. “Automated Depression Analysis Using Convolutional Neural Networks from Speech”. In: *Journal of Biomedical Informatics* 83 (2018), pp. 103–111. DOI: 10.1016/j.jbi.2018.05.007.
- [Hel13] Brian S. Helfer, Thomas F. Quatieri, James R. Williamson, Daryush D. Mehta, Rachelle Horwitz, and Bea Yu. “Classification of Depression State Based on Articulatory Precision”. In: *Interspeech 2013*. ISCA, 2013, pp. 2172–2176. DOI: 10.21437/Interspeech.2013-513.

- [Het08] S.E. Hetrick, A.G. Parker, I.B. Hickie, R. Purcell, A.R. Yung, and P.D. McGorry. “Early Identification and Intervention in Depressive Disorders: Towards a Clinical Staging Model”. In: *Psychotherapy and Psychosomatics* 77.5 (2008), pp. 263–270. doi: 10.1159/000140085.
- [Hir21] Daniel J. Hirst and Céline de Looze. “Measuring Speech. Fundamental Frequency and Pitch.” In: *Cambridge Handbook of Phonetics*. Ed. by Rachael-Anne Knight and Jane Setter. 1. Cambridge University Press, 2021, pp. 336–361.
- [Hol11] John H. Hollman, Eric M. McDade, and Ronald C. Petersen. “Normative Spatiotemporal Gait Parameters in Older Adults”. In: *Gait & Posture* 34.1 (2011), pp. 111–118. doi: 10.1016/j.gaitpost.2011.03.024. (Visited on 06/01/2025).
- [Hol80] Harry Hollien. “Vocal Indicators of Psychological Stress”. In: *Annals of the New York Academy of Sciences* 347.1 (1980), pp. 47–72. doi: 10.1111/j.1749-6632.1980.tb21255.x.
- [Hon08] Kiyoshi Honda. “Physiological Processes of Speech Production”. In: *Springer Handbook of Speech Processing*. Ed. by Jacob Benesty, M. Mohan Sondhi, and Yiteng Arden Huang. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 7–26. ISBN: 978-3-540-49125-5.
- [Hos19] Tahera Hossain, Md Shafiqul Islam, Md Atiqur Rahman Ahad, and Sozo Inoue. “Human Activity Recognition Using Earable Device”. 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*. London United Kingdom: ACM, 2019, pp. 81–84. ISBN: 978-1-4503-6869-8. doi: 10.1145/3341162.3343822.
- [Hwa18] Tong-Hun Hwang, Julia Reh, Alfred O. Effenberg, and Holger Blume. “Real-Time Gait Analysis Using a Single Head-Worn Inertial Measurement Unit”. In: *IEEE Transactions on Consumer Electronics* 64.2 (2018), pp. 240–248. doi: 10.1109/TCE.2018.2843289.
- [Hwa21] Tong-Hun Hwang, Julia Reh, Alfred O. Effenberg, and Holger Blume. “Validation of Real Time Gait Analysis Using a Single Head-Worn IMU”. In: *EKC 2019 Conference Proceedings*. Ed. by Jong Mun Park and Dong Ryeol Whang. Singapore: Springer, 2021, pp. 87–97. ISBN: 978-981-15-8350-6. doi: 10.1007/978-981-15-8350-6_8.

- [Jac19] Nicholas C. Jacobson, Hilary Weingarden, and Sabine Wilhelm. “Digital Biomarkers of Mood Disorders and Symptom Change”. In: *npj Digital Medicine* 2.1 (2019), pp. 1–3. doi: 10.1038/s41746-019-0078-0.
- [Jia15] Haomiao Jia, Matthew M. Zack, William W. Thompson, Alex E. Crosby, and Irving I. Gottesman. “Impact of Depression on Quality-Adjusted Life Expectancy (QALE) Directly as well as Indirectly Through Suicide”. In: *Social Psychiatry and Psychiatric Epidemiology* 50.6 (2015), pp. 939–949. doi: 10.1007/s00127-015-1019-0.
- [Jia18] Haihua Jiang, Bin Hu, Zhenyu Liu, Gang Wang, Lan Zhang, Xiaoyu Li, and Huanyu Kang. “Detecting Depression Using an Ensemble Logistic Regression Model Based on Multiple Speech Features”. In: *Computational and Mathematical Methods in Medicine* 2018.1 (2018), p. 6508319. doi: 10.1155/2018/6508319.
- [Jun22] Dawoon Jung, Jinwook Kim, and Kyung-Ryoul Mun. “Identifying Depression in the Elderly Using Gait Accelerometry”. In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Glasgow, Scotland, United Kingdom: IEEE, 2022, pp. 4946–4949. ISBN: 978-1-7281-2782-8. doi: 10.1109/EMBC48229.2022.9871877.
- [Jun23] Chang Keun Jung, Jinkyuk Kim, and Hye Chang Rhim. “Validation of an Ear-Worn Wearable Gait Analysis Device”. In: *Sensors* 23.3 (2023), p. 1244. doi: 10.3390/s23031244.
- [Ken98] Raymond D. Kent. *The Speech Sciences*. 2. print. San Diego, Calif.: Singular Publ. Group, 1998. ISBN: 978-1-56593-689-8.
- [Kit16] Naoki Kitagawa and Naomichi Ogiwara. “Estimation of Foot Trajectory During Human Walking by a Wearable Inertial Measurement Unit Mounted to the Foot”. In: *Gait & Posture* 45 (2016), pp. 110–114. doi: 10.1016/j.gaitpost.2016.01.014.
- [Kle02] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic Regression*. Springer, 2002.
- [Kon16] Weisheng Kong, Salvatore Sessa, Massimiliano Zecca, and Atsuo Takanishi. “Anatomical Calibration through Post-Processing of Standard Motion Tests Data”. In: *Sensors* 16.12 (2016). Number: 12 Publisher: Multidisciplinary Digital Publishing Institute, p. 2011. doi: 10.3390/s16122011.

- [Kre12] Jody Kreiman and Bruce R. Gerratt. “Perceptual Interaction of the Harmonic Source and Noise in Voice”. In: *The Journal of the Acoustical Society of America* 131.1 (2012), pp. 492–500. doi: 10.1121/1.3665997.
- [Kro09] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B. W. Williams, Joyce T. Berry, and Ali H. Mokdad. “The PHQ-8 as a Measure of Current Depression in the General Population”. In: *Journal of Affective Disorders* 114.1 (2009), pp. 163–173. doi: 10.1016/j.jad.2008.06.026.
- [Küd24] Arne Küderle, Martin Ullrich, Nils Roth, Malte Ollenschläger, Alzhraa A. Ibrahim, Hamid Moradi, Robert Richer, Ann-Kristin Seifer, Matthias Zürl, Raul C. Sîmpetru, Liv Herzer, Dominik Prossel, Felix Kluge, and Bjoern M. Eskofier. “Gaitmap—An Open Ecosystem for IMU-Based Human Gait Analysis and Algorithm Benchmarking”. In: *IEEE Open Journal of Engineering in Medicine and Biology* 5 (2024), pp. 163–172. doi: 10.1109/OJEMB.2024.3356791.
- [Kup16] Aleksandra Kupferberg, Lucy Bicks, and Gregor Hasler. “Social Functioning in Major Depressive Disorder”. In: *Neuroscience & Biobehavioral Reviews* 69 (2016), pp. 313–332. doi: 10.1016/j.neubiorev.2016.07.002.
- [Lea24] Samara Soares Leal, Stavros Ntalampiras, and Roberto Sassi. “Speech-Based Depression Assessment: A Comprehensive Survey”. In: *IEEE Transactions on Affective Computing* (2024). Conference Name: IEEE Transactions on Affective Computing, pp. 1–16. doi: 10.1109/TAFFC.2024.3521327.
- [Lem00] Matthias R Lemke, Thomas Wendorff, Brigitt Mieth, Katharina Buhl, and Martin Linnemann. “Spatiotemporal Gait Patterns During Over Ground Locomotion in Major Depression Compared with Healthy Controls”. In: *Journal of Psychiatric Research* 34.4 (2000), pp. 277–283. doi: 10.1016/S0022-3956(00)00017-0.
- [Ler08] Debra Lerner and Rachel Mosher Henke. “What Does Research Tell Us About Depression, Job Performance, and Work Productivity?” In: *Journal of Occupational and Environmental Medicine* 50.4 (2008), p. 401. doi: 10.1097/JOM.0b013e31816bae50.
- [Lev60] Howard Levene. “Robust Tests for Equality of Variances”. In: *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Ed. by Ingram Olkin, Harold Hotelling, et al. Stanford University Press, 1960, pp. 278–292. isbn: 0-8047-0596-8.

- [Li17] Gang Li, Zhe Zhang, and Guoxing Wang. “Emotion Recognition Based on Low-Cost In-Ear EEG”. In: *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. Torino: IEEE, 2017, pp. 1–4. ISBN: 978-1-5090-5803-7. DOI: 10.1109/BIOCAS.2017.8325198.
- [Li21] Wentao Li, Qingxiang Wang, Xin Liu, and Yanhong Yu. “Simple Action for Depression Detection: Using Kinect-Recorded Human Kinematic Skeletal Data”. In: *BMC Psychiatry* 21.1 (2021), p. 205. DOI: 10.1186/s12888-021-03184-4.
- [Li25] Jian Li, Yuliang Zhao, Huawei Zhang, Wayne Jason Li, Changzeng Fu, Chao Lian, and Peng Shan. “Image Encoding and Fusion of Multi-Modal Data Enhance Depression Diagnosis in Parkinson’s Disease Patients”. In: *IEEE Transactions on Affective Computing* 16.1 (2025), pp. 145–160. DOI: 10.1109/TAFFC.2024.3418415.
- [Low20] Daniel M. Low, Kate H. Bentley, and Satrajit S. Ghosh. “Automated Assessment of Psychiatric Disorders Using Speech: A Systematic Review”. In: *Laryngoscope Investigative Otolaryngology* 5.1 (2020), pp. 96–116. DOI: 10.1002/lto2.354.
- [Lu21] Haifeng Lu, Wei Shao, Edith Ngai, Xiping Hu, and Bin Hu. “A New Skeletal Representation Based on Gait for Depression Detection”. In: *2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM)*. Shenzhen, China: IEEE, 2021, pp. 1–6. ISBN: 978-1-7281-6267-6. DOI: 10.1109/HEALTHCOM49281.2021.9399002.
- [Lun17] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.
- [Mam09] Alessia Mammone, Marco Turchi, and Nello Cristianini. “Support Vector Machines”. In: *WIREs Computational Statistics* 1.3 (2009), pp. 283–289. DOI: 10.1002/wics.49.
- [Man47] H. B. Mann and D. R. Whitney. “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other”. In: *The Annals of Mathematical Statistics* 18.1 (1947), pp. 50–60. DOI: 10.1214/aoms/1177730491.
- [Mej21] Yohanna MejiaCruz, Jean Franco, Garrett Hainline, Stacy Fritz, Zhaoshuo Jiang, Juan M. Caicedo, Benjamin Davis, and Victor Hirth. “Walking Speed Measurement Technology: a Review”. In: *Current Geriatrics Reports* 10.1 (2021), pp. 32–41. DOI: 10.1007/s13670-020-00349-z.

- [Min18] Chulhong Min, Akhil Mathur, and Fahim Kawsar. “Exploring Sudio and Kinetic Sensing on Earable Devices”. In: *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*. Munich Germany: ACM, 2018, pp. 5–10. ISBN: 978-1-4503-5842-2. DOI: 10.1145/3211960.3211970.
- [Moh17] David C. Mohr, Mi Zhang, and Stephen M. Schueller. “Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning”. In: *Annual Review of Clinical Psychology* 13.1 (2017), pp. 23–47. DOI: 10.1146/annurev-clinpsy-032816-044949.
- [Moo16] Yaejin Moon, JongHun Sung, Ruopeng An, Manuel E. Hernandez, and Jacob J. Sosnoff. “Gait Variability in People with Neurological Disorders: A Systematic Review and Meta-Analysis”. In: *Human Movement Science* 47 (2016), pp. 197–208. DOI: 10.1016/j.humov.2016.03.010.
- [Mun12] James C. Mundt, Adam P. Vogel, Douglas E. Feltner, and William R. Lenderking. “Vocal Acoustic Biomarkers of Depression Severity and Treatment Response”. In: *Biological Psychiatry. Novel Pharmacotherapies for Depression* 72.7 (2012), pp. 580–587. DOI: 10.1016/j.biopsych.2012.03.015.
- [Nar20] Nematjon Narziev, Hwarang Goh, Kobiljon Toshnazarov, Seung Ah Lee, Kyong-Mee Chung, and Youngtae Noh. “STDD: Short-Term Depression Detection with Passive Sensing”. In: *Sensors* 20.5 (2020), p. 1396. DOI: 10.3390/s20051396.
- [Nas16] Md Nasir, Arindam Jati, Prashanth Gurunath Shivakumar, Sandeep Nallan Chakravarthula, and Panayiotis Georgiou. “Multimodal and Multiresolution Depression Detection from Speech and Facial Landmark Features”. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. AVEC ’16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 43–50. ISBN: 978-1-4503-4516-3. DOI: 10.1145/2988257.2988261.
- [OBr17] J. T. O’Brien, P. Gallagher, D. Stow, N. Hammerla, T. Ploetz, M. Firbank, C. Ladha, K. Ladha, D. Jackson, R. McNaney, I. N. Ferrier, and P. Olivier. “A Study of Wrist-Worn Activity Measurement as a Potential Real-World Biomarker for Late-Life Depression”. In: *Psychological Medicine* 47.1 (2017), pp. 93–102. DOI: 10.1017/S0033291716002166.
- [Oth21] Alice Othmani, Daoud Kadoch, Kamil Bentounes, Emna Rejaibi, Romain Alfred, and Abdenour Hadid. “Towards Robust Deep Neural Networks for Affect and Depression Recognition from Speech”. In: *Pattern Recognition. ICPR International Workshops*

- and Challenges*. Ed. by Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani. Cham: Springer International Publishing, 2021, pp. 5–19. ISBN: 978-3-030-68790-8. DOI: 10.1007/978-3-030-68790-8_1.
- [Par17] Pallavi R. Parekh and M. M. Patil. “Clinical Depression Detection for Adolescent by Speech Features”. In: *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*. 2017, pp. 3453–3457. DOI: 10.1109/ICECDS.2017.8390102.
- [Ped11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [Per10] Jacquelin Perry. *Gait Analysis: Normal and Pathological Function, Second Edition*. 2nd ed. Thorofare: SLACK, Incorporated, 2010. ISBN: 978-1-55642-766-4.
- [Pet09] Leif E. Peterson. “K-Nearest Neighbor”. In: *Scholarpedia* 4.2 (2009), p. 1883. DOI: 10.4249/scholarpedia.1883.
- [Pfo97] Bruce Pfohl, Nancee Blum, and Mark Zimmerman. *Structured Interview for DSM-IV Personality: Sidp-IV*. American Psychiatric Pub, 1997.
- [Pla23] Alexis Plaquet and Hervé Bredin. “Powerset Multi-Class Cross Entropy Loss for Neural Speaker Diarization”. In: (2023). DOI: 10.48550/ARXIV.2310.13025.
- [Pra20] Jay Prakash, Zhijian Yang, Yu-Lin Wei, and Romit Roy Choudhury. “STEAR: Robust Step Counting from Earables”. In: *Proceedings of the 1st International Workshop on Earable Computing*. EarComp’19. New York, NY, USA: Association for Computing Machinery, 2020, pp. 36–41. ISBN: 978-1-4503-6902-2. DOI: 10.1145/3345615.3361133.
- [Rad14] Saša Radovanović, Milica Jovičić, Nadja P. Marić, and Vladimir Kostić. “Gait Characteristics in Patients with Major Depression Performing Cognitive and Motor Tasks While Walking”. In: *Psychiatry Research* 217.1 (2014), pp. 39–46. DOI: 10.1016/j.psychres.2014.02.001.

- [Ram11] Lorraine Ramig, Cynthia Fox, and Shimon Sapir. “Speech and Voice Disorders in Parkinson’s Disease”. In: *Parkinson’s Disease*. Ed. by C. Warren Olanow, Fabrizio Stocchi, and Anthony E. Lang. 1st ed. Wiley, 2011, pp. 346–360. ISBN: 978-1-4051-9185-2.
- [Ram21] Dinesh Ramoo. *Psychology of Language*. BCcampus, 2021. ISBN: 978-1-77420-132-9.
- [Ram98] Lorraine Olson Ramig and Katherine Verdolini. “Treatment Efficacy: Voice Disorders”. In: *Journal of Speech, Language, and Hearing Research* 41.1 (1998). doi: 10.1044/jslhr.4101.s101.
- [Rej22] Emna Rejaibi, Ali Komaty, Fabrice Meriaudeau, Said Agrebi, and Alice Othmani. “MFCC-Based Recurrent Neural Network for Automatic Clinical Depression Recognition and Assessment from Speech”. In: *Biomedical Signal Processing and Control* 71 (2022), p. 103107. doi: 10.1016/j.bspc.2021.103107.
- [Ric21] Robert Richer, Arne Küderle, Martin Ullrich, Nicolas Rohleder, and Bjoern Eskofier. “BioPsyKit: A Python Package for the Analysis of Biopsychological Data”. In: *Journal of Open Source Software* 6.66 (2021), p. 3702. doi: 10.21105/joss.03702.
- [Ric24] Robert Richer, Veronika Koch, Luca Abel, Felicitas Hauck, Miriam Kurz, Veronika Ringgold, Victoria Müller, Arne Küderle, Lena Schindler-Gmelch, Bjoern M. Eskofier, and Nicolas Rohleder. “Machine Learning-Based Detection of Acute Psychosocial Stress from Body Posture and Movements”. In: *Scientific Reports* 14.1 (2024), p. 8251. doi: 10.1038/s41598-024-59043-1.
- [Röd22] Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. “Sensing with Earables: A Systematic Literature Review and Taxonomy of Phenomena”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6.3 (2022), 135:1–135:57. doi: 10.1145/3550314.
- [Rot90] Kenneth J. Rothman. “No Adjustments Are Needed for Multiple Comparisons”. In: *Epidemiology* 1.1 (1990), p. 43.
- [Roy03] Nelson Roy. “Functional Dysphonia”. In: *Current Opinion in Otolaryngology & Head & Neck Surgery* 11.3 (2003), pp. 144–148.
- [Ryk21] Yuri Rykov, Thuan-Quoc Thach, Iva Bojic, George Christopoulos, and Josip Car. “Digital Biomarkers for Depression Screening with Wearable Devices: Cross-sectional Study With Machine Learning Modeling”. In: *JMIR mHealth and uHealth* 9.10 (2021), e24872. doi: 10.2196/24872.

- [San08] Jan van Santen, Taniya Mishra, and Esther Klabbers. “Prosodic Processing”. In: *Springer Handbook of Speech Processing*. Ed. by Jacob Benesty, M. Mohan Sondhi, and Yiteng Arden Huang. Berlin, Heidelberg: Springer, 2008, pp. 471–488. ISBN: 978-3-540-49127-9.
- [San15] Ziggi Ivan Santini, Ai Koyanagi, Stefanos Tyrovolas, Catherine Mason, and Josep Maria Haro. “The Association Between Social Relationships and Depression: A Systematic Review”. In: *Journal of Affective Disorders* 175 (2015), pp. 53–65. doi: 10.1016/j.jad.2014.12.049.
- [Sch13] Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency. “Investigating Voice Quality as a Speaker-Independent Indicator of Depression and PTSD”. In: *Interspeech 2013*. ISCA, 2013, pp. 847–851. doi: 10.21437/Interspeech.2013-240.
- [Sch16] Sandra Schwab and Jean-Philippe Goldman. “Do Speakers Show Different F0 When They Speak in Different Languages? The Case of English, French and German?” In: (2016). doi: 10.5167/UZH-126023.
- [Sei23] Ann-Kristin Seifer, Eva Dorschky, Arne Küderle, Hamid Moradi, Ronny Hannemann, and Björn M. Eskofier. “EarGait: Estimation of Temporal Gait Parameters from Hearing Aid Integrated Inertial Sensors”. In: *Sensors* 23.14 (2023), p. 6565. doi: 10.3390/s23146565.
- [Sei24] Ann-Kristin Seifer, Arne Küderle, Eva Dorschky, Hamid Moradi, Ronny Hannemann, and Bjoern M. Eskofier. “Step Length and Gait Speed Estimation Using a Hearing Aid Integrated Accelerometer: A Comparison of Different Algorithms”. In: *IEEE Journal of Biomedical and Health Informatics* 28.11 (2024), pp. 6619–6628. doi: 10.1109/JBHI.2024.3454824.
- [Sha21] Rahul K. Sharma, Alexander Chern, and Justin S. Golub. “Age-Related Hearing Loss and the Development of Cognitive Impairment and Late-Life Depression: A Scoping Overview”. In: *Seminars in Hearing* 42.1 (2021), pp. 10–25. doi: 10.1055/s-0041-1725997.
- [Sha22] Wei Shao, Zhiyang You, Lesheng Liang, Xiping Hu, Chengming Li, Wei Wang, and Bin Hu. “A Multi-Modal Gait Analysis-Based Detection System of the Risk of Depression”. In: *IEEE Journal of Biomedical and Health Informatics* 26.10 (2022), pp. 4859–4868. doi: 10.1109/JBHI.2021.3122299.

- [Sha65] S. S. Shapiro and M. B. Wilk. “An Analysis of Variance Test for Normality (Complete Samples)”. In: *Biometrika* 52.3/4 (1965), pp. 591–611. doi: 10.2307/2333709.
- [Shi21] Daun Shin, Won Ik Cho, C. Hyung Keun Park, Sang Jin Rhee, Min Ji Kim, Hyunju Lee, Nam Soo Kim, and Yong Min Ahn. “Detection of Minor and Major Depression through Voice as a Biomarker Using Machine Learning”. In: *Journal of Clinical Medicine* 10.14 (2021), p. 3046. doi: 10.3390/jcm10143046.
- [Sir20] Sirakorn. *An Illustration for the Concept of Bootstrap Aggregating*. 2020. URL: <https://commons.wikimedia.org/w/index.php?curid=85888768> (visited on 05/22/2025).
- [Sob97] C. Sobin and H. A. Sackeim. “Psychomotor Symptoms of Depression”. In: *The American Journal of Psychiatry* 154.1 (1997), pp. 4–17. doi: 10.1176/ajp.154.1.4.
- [Stö15] Tino Stöckel, Robert Jacksteit, Martin Behrens, Ralf Skripitz, Rainer Bader, and Anett Mau-Moeller. “The Mental Representation of the Human Gait in Young and Older Adults”. In: *Frontiers in Psychology* 6 (2015). doi: 10.3389/fpsyg.2015.00943.
- [Sto98] H Stolze, J. P Kuhtz-Buschbeck, C Mondwurf, K Jöhnk, and L Friege. “Retest Reliability of Spatiotemporal Gait Parameters in Children and Adults”. In: *Gait & Posture* 7.2 (1998), pp. 125–130. doi: 10.1016/S0966-6362(97)00043-X.
- [Stu08] Student. “The Probable Error of a Mean”. In: *Biometrika* 6.1 (1908). Publisher: [Oxford University Press, Biometrika Trust], pp. 1–25. doi: 10.2307/2331554.
- [Sup23] Sukit Suparatpinyo and Nuanwan Soonthornphisaj. “Smart Voice Recognition Based on Deep Learning for Depression Diagnosis”. In: *Artificial Life and Robotics* 28.2 (2023), pp. 332–342. doi: 10.1007/s10015-023-00852-4.
- [Tag95] Gentaro Taga. “A Model of the Neuro-Musculo-Skeletal System for Human Locomotion”. In: *Biological Cybernetics* 73.2 (1995), pp. 97–111. doi: 10.1007/BF00204048.
- [Tia23] Han Tian, Zhang Zhu, and Xu Jing. “Deep Learning for Depression Recognition from Speech”. In: *Mobile Networks and Applications* (2023). doi: 10.1007/s11036-022-02086-3.

- [Tot21] Ermal Toto, ML Tlachac, and Elke A. Rundensteiner. “AudiBERT: A Deep Transfer Learning Multimodal Classification Framework for Depression Screening”. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. CIKM ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 4145–4154. ISBN: 978-1-4503-8446-9. DOI: 10.1145/3459637.3481895.
- [Uça20] Muhammed Kürşad Uçar, Majid Nour, Hatem Sindi, and Kemal Polat. “The Effect of Training and Testing Process on Machine Learning in Biomedical Datasets”. In: *Mathematical Problems in Engineering 2020* (2020), pp. 1–17. DOI: 10.1155/2020/2836236.
- [Val13] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. “AVEC 2013: the continuous audio/visual emotion and depression recognition challenge”. In: *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. AVEC ’13. New York, NY, USA: Association for Computing Machinery, 2013, pp. 3–10. ISBN: 978-1-4503-2395-6. DOI: 10.1145/2512530.2512533.
- [Val16] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. “AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge”. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. AVEC ’16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 3–10. ISBN: 978-1-4503-4516-3. DOI: 10.1145/2988257.2988258.
- [Vas17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.
- [Váz20] Adrián Vázquez-Romero and Ascensión Gallardo-Antolín. “Automatic Detection of Depression in Speech Using Ensemble Convolutional Neural Networks”. In: *Entropy* 22.6 (2020), p. 688. DOI: 10.3390/e22060688.
- [Ver21] Dhruv Verma, Sejal Bhalla, Dhruv Sahnan, Jainendra Shukla, and Aman Parnami. “ExpressEar: Sensing Fine-Grained Facial Expressions with Earables”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5.3 (2021), 129:1–129:28. DOI: 10.1145/3478085.

- [Wag23] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. “Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.9 (2023), pp. 10745–10759. doi: 10.1109/TPAMI.2023.3263585.
- [Wan07] Philip S Wang, Matthias Angermeyer, Guilherme Borges, Ronny Bruffaerts, Wai Tat Chiu, Giovanni de Girolamo, John Fayyad, Oye Gureje, Josep Maria Haro, Yueqin Huang, Ronald C Kessler, Viviane Kovess, Daphna Levinson, Yoshibumi Nakane, Mark A Oakley Browne, Johan H Ormel, José Posada-Ville, Sergio Aguilar-Gaxiola, Jordi Alonso, Sing Lee, Steven Heeringa, Beth-Ellen Pennell, Somnath Chatterji, and T. Bedirhan Üstün. “Delay and Failure in Treatment Seeking After First Onset of Mental Disorders in the Eorld Health Organization’s World Mental Health Survey Initiative”. In: *World Psychiatry* 6.3 (2007), pp. 177–185.
- [Wan21] Yameng Wang, Jingying Wang, Xiaoqian Liu, and Tingshao Zhu. “Detecting Depression Through Gait Data: Examining the Contribution of Gait Features in Recognizing Depression”. In: *Frontiers in Psychiatry* 12 (2021). doi: 10.3389/fpsy.2021.661213.
- [Wan22] Tiffany V Wang and Phillip C Song. “Neurological Voice Disorders: A Review”. In: *International Journal of Head and Neck Surgery* 13.1 (2022), pp. 32–40. doi: 10.5005/jp-journals-10001-1521.
- [Was19] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. “Moving to a World Beyond “ $p < 0.05$ ””. In: *The American Statistician* 73.sup1 (2019), pp. 1–19. doi: 10.1080/00031305.2019.1583913.
- [Wel47] B. L. Welch. “The Generalization of ‘Student’s’ Problem When Several Different Population Variances are Involved”. In: *Biometrika* 34.1-2 (1947), pp. 28–35. doi: 10.1093/biomet/34.1-2.28.
- [WHO04] WHO. *ICD-10 : International Statistical Classification of Diseases and Related Health Problems : Tenth Revision*. World Health Organization, 2004. ISBN: 978-92-4-154649-2.
- [WHO23] WHO. *Depressive disorder (depression)*. 2023. URL: <https://www.who.int/news-room/fact-sheets/detail/depression> (visited on 09/04/2024).

- [Wu23a] Pingping Wu, Ruihao Wang, Han Lin, Fanlong Zhang, Juan Tu, and Miao Sun. “Automatic Depression Recognition by Intelligent Speech Signal Processing: A Systematic Survey”. In: *CAAI Transactions on Intelligence Technology* 8.3 (2023), pp. 701–711. doi: 10.1049/cit2.12113.
- [Wu23b] Wen Wu, Chao Zhang, and Philip C. Woodland. “Self-Supervised Representations in Speech-Based Depression Detection”. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10094910.
- [Yam20] Mao Yamamoto, Akihiro Takamiya, Kyosuke Sawada, Michitaka Yoshimura, Momoko Kitazawa, Kuo-ching Liang, Takanori Fujita, Masaru Mimura, and Taishiro Kishimoto. “Using Speech Recognition Technology to Investigate The Association Between Timing-Related Speech Features and Depression Severity”. In: *PLOS ONE* 15.9 (2020), e0238726. doi: 10.1371/journal.pone.0238726.
- [Yan17] Le Yang, Dongmei Jiang, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. “Multimodal Measurement of Depression Using Deep Learning Models”. In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. AVEC '17*. New York, NY, USA: Association for Computing Machinery, 2017, pp. 53–59. ISBN: 978-1-4503-5502-5. doi: 10.1145/3133944.3133948.
- [Yus16] Salifu Yusif, Jeffrey Soar, and Abdul Hafeez-Baig. “Older People, Assistive Technologies, and the Barriers to Adoption: A Systematic Review”. In: *International Journal of Medical Informatics* 94 (2016), pp. 112–116. doi: 10.1016/j.ijmedinf.2016.07.004.
- [Zha16] Zhaoyan Zhang. “Mechanics of Human Voice Production and Control”. In: *The Journal of the Acoustical Society of America* 140.4 (2016), pp. 2614–2635. doi: 10.1121/1.4964509.
- [Zha19] Nan Zhao, Zhan Zhang, Yameng Wang, Jingying Wang, Baobin Li, Tingshao Zhu, and Yuanyuan Xiang. “See Your Mental State from Your Walk: Recognizing Anxiety and Depression Through Kinect-Recorded Gait Data”. In: *PLOS ONE* 14.5 (2019), e0216591. doi: 10.1371/journal.pone.0216591.
- [Zha21] Pingyue Zhang, Mengyue Wu, Heinrich Dinkel, and Kai Yu. “DEPA: Self-Supervised Audio Embedding for Depression Detection”. In: *Proceedings of the 29th ACM International Conference on Multimedia. MM '21*. New York, NY, USA: Association

for Computing Machinery, 2021, pp. 135–143. ISBN: 978-1-4503-8651-7. DOI: 10.1145/3474085.3479236.

Appendix A

Study Protocol

dEARpression: Erkennung von depressiven Symptomen anhand verschiedener Sprach- und Bewegungsparameter – Eine Pilotstudie

Studienprotokoll

eines sonstigen biomedizinischen Forschungsvorhabens

des Lehrstuhls für Maschinelles Lernen und Datenanalytik (MaD Lab)

Versionsnummer	1
Versionsdatum	17.01.2025
Autoren	Ann-Kristin Seifer (M.Sc), Robert Richer (M.Sc), Lydia Rupp (M.Sc.), Prof. Björn Eskofier, Prof. Matthias Berking

Inhalt

1. Verantwortlichkeiten und Kontaktdaten	4
2. Zusammenfassung	4
3. Wissenschaftlicher Hintergrund.....	5
4. Studienziele	6
4.1. Zielgrößen	6
5. Studiendesign.....	7
5.1. Monozentrisch/multizentrisch.....	7
5.2. Intervention / Kontrolle	7
5.3. Randomisierung	7
5.4. Verblindung.....	7
6. Studienpopulation (Kollektiv).....	7
6.1. Ein- und Ausschlusskriterien	7
6.2. Anzahl.....	7
6.3. Explorative Studie: Erläuterung zur statistischen Methodik, Begründung der gewählten Fallzahl.....	7
6.4. Rekrutierungswege und -maßnahmen.....	8
7. Studienablauf	8
7.1. Verfahren zur Aufklärung und Einholung der Einwilligung	8
7.2. Sensorik.....	8
7.3. Maßnahmen (Intervention/Kontrolle)	8
7.4. Zeitlicher Ablauf (Termine) und Studiendauer für den/die einzelne:n Proband:in	10
7.5. Gesamtdauer der Studie	10
8. Nutzen-Risiko-Abwägung (Sicherheitsaspekte)	10
8.1. Studienbedingte Risiken.....	10
8.2. Mit der Studie verbundener Nutzen	11
8.3. Abbruchkriterien	11
8.4. Statement zur ärztlichen Vertretbarkeit	11
9. Datenmanagement und Datenschutz	11
9.1. Datenerfassung, -speicherung	11
9.2. Anonymisiert/pseudonymisiert	12
9.3. Datenweitergabe.....	12
9.4. Widerruf und Datenlöschung.....	12

10. Umgang mit Biomaterialien	12
11. Referenzen	13
12. Unterschriften: Studienleiter:in (Antragsteller:in)	14

1. Verantwortlichkeiten und Kontaktdaten

Studienleitung	Prof. Dr. Björn Eskofier Lehrstuhl für Maschinelles Lernen und Datenanalytik (MaD Lab) Department Artificial Intelligence for Biomedical Engineering Friedrich-Alexander-Universität Erlangen-Nürnberg D-91052 Erlangen
Stellvertretende Studienleitung	Prof. Dr. Matthias Berking Lehrstuhl für Klinische Psychologie und Psychotherapie (KliPs) Institut für Psychologie Friedrich-Alexander-Universität Erlangen-Nürnberg D-91052 Erlangen
Beteiligte Personen	Robert Richer (MaD Lab) Ann-Kristin Seifer (MaD Lab) Lydia Rupp (KliPs) Lena Gmelch (KliPs)
Kooperationspartner	Sivantos GmbH Henri-Dunant-Str. 100 D-91058 Erlangen

2. Zusammenfassung

Titel	dEARpression: Erkennung von depressiven Symptomen anhand verschiedener Sprach- und Bewegungsparameter – Eine Pilotstudie
Studiendesign	Eine monozentrische Studie bestehend aus einer offenen, prospektiven Evaluationsstudie
Rationale (kurz)	Ziel dieses Projektes ist es, mit Hilfe der Sensoren, die in einem herkömmlichen Hörgerät integriert sind (Accelerometer, Mikrofon), depressive Symptome und Depressionen zu erkennen. Hörgeräte bieten den Vorteil, dass sie bereits im Alltag einiger Menschen integriert sind und täglich über einen langen Zeitraum getragen werden. Die Erkennung basiert sowohl auf der Analyse verschiedener Stimm-Charakteristiken als auch auf Daten über den Gang und die Bewegung der Teilnehmer:innen. Beides wird über Sensoren im Hörgerät aufgezeichnet, welche den Proband:innen angelegt werden. Anschließend werden diese Daten von einem hierzu entwickelten Algorithmus ausgewertet, der bei der Einordnung depressiver Symptome unterstützen soll. Der Depressionsgrad der Teilnehmer:innen wird mit Hilfe eines standardisierten Fragebogens

	ermittelt. Es handelt sich hierbei um eine Pilotstudie, um erste Einblicke in die Machbarkeit und Datenqualität mit Hilfe von realistischen Daten zu bekommen.
Studienziele	• Stimm- und Bewegungsdaten über Hörgeräte aufzeichnen und analysieren • Bewegungsdaten über Hörgeräte erfassen • Depressive Symptome anhand der Sprach- und Bewegungsdaten analysieren und einordnen
Teilnehmerzahl	20 Proband:innen - 10 mit Depression (PHQ-Score ≥ 10 [1], Fallgruppe) - 10 Gesunde (Kontrollgruppe)
Ein- und Ausschluss Kriterien	Siehe Punkt 6

3. Wissenschaftlicher Hintergrund

Der Lehrstuhl für Maschinelles Lernen und Datenanalytik (MaD) am Department Artificial Intelligence in Biomedical Engineering (AIBE) beschäftigt sich mit den Anwendungen von Mustererkennungsalgorithmen in den Bereichen Sport, Bewegung, Gesundheitsweisen, etc. Die Fragestellung, die von Mitarbeitenden des Lehrstuhls für Maschinelles Lernen und Datenanalytik in dieser Studie untersucht werden soll, beinhaltet die Datenerhebung, Analyse und Auswertung gewonnener Sprachdaten (z. B. Berechnung von Stimmmerkmalen). Zusätzlich sollen Bewegungsdaten mit Hilfe von im Hörgerät integrierten Bewegungssensoren aufgezeichnet werden. Die Ergebnisse dieser Auswertung können einen wertvollen Beitrag zur Forschung im Bereich der Erkennung depressiver Symptome und Depressionen von Menschen mit einem alltagstauglichen Sensor leisten.

Depression ist eine der häufigsten psychischen Erkrankungen mit über 280 Millionen Betroffenen weltweit und stellt eine hohe Belastung für Erkrankte dar [2]. Eine frühzeitige Erkennung spielt eine große Rolle im Genesungsprozess mit höheren Remissionsraten [3].

Die Diagnose einer Depression erfolgt in der Regel durch ausgebildete Fachkräfte, beispielsweise Ärzte, Psychiater oder Psychologen. Dabei finden strukturierte Interviews und standardisierte Fragebögen Anwendung, beispielsweise die Hamilton Rating Scale for Depression (HDRS), das Beck Depression Inventory-II (BDI-II) sowie der Patient Health Questionnaire (PHQ). Diese Beurteilungen setzen jedoch voraus, dass die Patienten medizinische Hilfe bei den genannten Fachkräften suchen.

Ein sich etablierendes Forschungsgebiet befasst sich mit alternativen Ansätzen zur Erkennung von Depressionen. Dabei wird das Potenzial objektiver digitaler Biomarker wie Stimme, Aktivität, Gangart, Schlafdauer und weiterer Faktoren bewertet [4]. Ein häufig verwendeter Biomarker für die Erkennung von Depressionen ist die Stimme, da Personen mit Depressionen bemerkenswerte Veränderungen in ihren Sprachmustern aufweisen (vgl. hierzu auch [5], [6]). Dazu gehören beispielsweise eine geringere Intensität, eine Abnahme der Grundfrequenzabweichung oder des Tonhöhenbereichs sowie eine Verlangsamung der

Sprechgeschwindigkeit [7]. Die Anwendung maschinellen Lernens hat die Entwicklung zahlreicher automatischer Sprachanalyseverfahren ermöglicht, welche paralinguistische Merkmale nutzen [5]. Ein weiterer objektiver Biomarker für Menschen mit Depressionen ist der Gang und andere damit zusammenhängende körperliche Merkmale, darunter eine zusammengesackte Haltung, ein gestörtes dynamisches Gleichgewicht und eine verringerte Geschwindigkeit [8].

Die Daten können auf verschiedene Weise gewonnen werden. Neben externen Sensoren, die in der Umgebung fixiert sind, existieren tragbare Sensorgeräte, welche vom Benutzer getragen werden können und bereits in dessen Alltag integriert sind. In diese Kategorie tragbarer Technologie lassen sich Smartphones, Smartwatches und Hörgeräte einordnen [9].

Obwohl die Möglichkeit bereits gezeigt wurde, Depressionen mithilfe von Stimmanalyse und Gangmustern zu diagnostizieren, wurde bislang keine Untersuchung darüber angestellt, ob sich die Kombination beider Informationen zur Verbesserung der Depressionserkennung eignet. Des Weiteren sind die meisten bestehenden Ansätze für den alltäglichen Gebrauch ungeeignet, da Benutzer:innen einen oder mehrere Sensoren anlegen muss. Ein Hörgerät hat den Vorteil, dass es Stimme, Gang und andere Bewegungen direkt aufzeichnen kann, da es sowohl einen Bewegungs- als auch einen Mikrofonsensor integriert hat. Während Earables im Allgemeinen immer beliebter werden [9], [10], stellt die Kombination von Sensoren am Ohr (oder Kopf) zur Erkennung von Depressionen einen neuartigen Ansatz dar, der in der Literatur noch nicht dokumentiert wurde.

4. Studienziele

Ziel dieses Projekts ist es, einen Algorithmus zur Klassifizierung einer möglichen Depression zu entwickeln und zu evaluieren. Die Daten dafür sollen mit Hilfe der Sensoren eines Hörgeräts aufgenommen werden. Das Studiendesign ist explorativ.

Zu den aufgenommenen Daten zählen Sprachdaten, wobei hier nicht der Inhalt, sondern das Sprachverhalten, d.h. paralinguistische Merkmale wie Stimmfarbe und Tonhöhe, analysiert werden soll. Zusätzlich werden Gang- und Bewegungsdaten erhoben und analysiert.

Insgesamt sollen die Fragen beantwortet werden, wie gut eine Depression mit Hilfe von Hörgerätsensoren klassifiziert werden kann und welche Merkmale einen statistisch signifikanten Einfluss auf das Ergebnis haben.

4.1. Zielgrößen

Die Zielgrößen werden aus den Daten und neu entwickelten und existierenden Algorithmen zur Erkennung einer möglichen depressiven Erkrankung gewonnen. Weitere Zielgrößen werden aus den Ergebnissen des Fragebogens (PHQ-8; umfassende Depressionsskala) ermittelt. Mit Hilfe der Zielgrößen werden Korrelationen der Daten mit den Ergebnissen des Fragebogens untersucht.

5. Studiendesign

5.1. Monozentrisch/multizentrisch

Monozentrische Studie

5.2. Intervention / Kontrolle

Es werden keine invasiven Maßnahmen durchgeführt.

5.3. Randomisierung

Eine Randomisierung über verschiedene Versuche findet statt.

5.4. Verblindung

Die Daten werden bei der Studienausswertung verblindet ausgewertet.

6. Studienpopulation (Kollektiv)

6.1. Ein- und Ausschlusskriterien

Die folgenden Einschlusskriterien sind vorgesehen:

- (1) Alter ≥ 18 Jahre
- (2) Muttersprache Deutsch
- (3) Fähigkeit selbstständig zu gehen
- (4) Bereitschaft und Fähigkeit teilzunehmen

Die folgenden Ausschlusskriterien sind vorgesehen:

- (1) Psychotische Störung
- (2) Akute Suchterkrankung
- (3) Bipolare Störung
- (4) Neurodegenerative oder muskuloskelettale Erkrankungen, Beeinträchtigung der Mobilität
- (5) Versorgter Hörverlust (Tragen eines Hörgeräts im Alltag)
- (6) Nicht-einwilligungsfähige Personen

Alle potenzielle Proband:innen werden durch uns aufgeklärt und werden gefragt, ob sie an der Studie teilnehmen möchten. Einer Teilnahme müssen sie ausdrücklich mit Unterschrift zustimmen. Die Teilnahme ist für alle Proband:innen freiwillig und kann jeder Zeit und ohne Angabe von Gründen abgebrochen werden.

6.2. Anzahl

Es werden 20 Proband:innen erwartet (Fallgruppe: 10 Depressive, Kontrollgruppe: 10 Gesunde).

6.3. Explorative Studie: Erläuterung zur statistischen Methodik, Begründung der gewählten Fallzahl

Die Fallzahl (20 Proband:innen) basiert auf ähnlichen Studien zur Entwicklung von maschinellen Lernalgorithmen.

6.4. Rekrutierungswege und -maßnahmen

Die Teilnehmer:innen werden in Zusammenarbeit mit dem Lehrstuhl für klinische Psychologie (KliPs) sowie durch eigene Initiativen wie Flyer und Aushänge, welche in Einrichtungen mit Menschen mit Depression ausgelegt werden, rekrutiert.

7. Studienablauf

7.1. Verfahren zur Aufklärung und Einholung der Einwilligung

Die Proband:innen werden über die geplante Studie vor der Studie in einem Gespräch und mittels des beiliegenden Aufklärungsbogens aufgeklärt und willigen schriftlich in die Studienteilnahme ein.

7.2. Sensorik

Bei den Sensoren handelt es sich um ein Mikrofon und einen Beschleunigungssensor, welche am Ohr und integriert in einem Hörgerätgehäuse getragen werden. Die Sensoren haben keinerlei Einfluss auf das Gehvermögen der Proband:innen. Die Verstärkung des Hörgeräts ist zu jeder Zeit deaktiviert und beeinflusst daher nicht das Hörvermögen der Proband:innen.

7.3. Maßnahmen (Intervention/Kontrolle)

Das folgende Vorgehen wird verwendet:

- Aufklärung der Studie durch Studienleiter:in oder einem/r Stellvertreter:in
- Lesen und Unterschreiben der Einwilligung- und Datenschutzerklärung
- Beantwortung eines standardisierten Fragebogens zur Erfassung des Vorliegens einer depressiven Symptomatik (Patient Health Questionnaire PHQ-8 [12], [13])

Die Studie besteht aus zwei verschiedenen Experimenten: (A) Aufnahme der Sprachdaten, sowie (B) Aufnahme der Bewegungsdaten (siehe hierzu auch Abbildung 1). Diese werden in randomisierter Reihenfolge stattfinden, um den Einfluss aufeinander zu minimieren.

A) Aufnahme der Sprachdaten

- Erklärung des Ablaufs
- Individuelle Anpassung der Hörerlänge und Hörgerät-Schirmchens (Dome)
- Anbringen der Sensorik am Ohr, je ein Gerät links und rechts
- Kopplung der Hörgeräte (per Kabel) mit einer Soundbox mit integrierten Speicherkarte
- Anbringung eines Stirnbands mit Sensoren zur Aufzeichnung von Bewegungsdaten während des Interviews
- Aufnahmen von verschiedenen Bewegungen (Stehen, Sit-To-Stand-Test) zur Kalibrierung der Sensoren zur Bewegungsaufzeichnung [14]

- Datenaufnahme während eines Interviews, welches darauf abzielt, möglichst viele paralinguistische Daten aufzunehmen

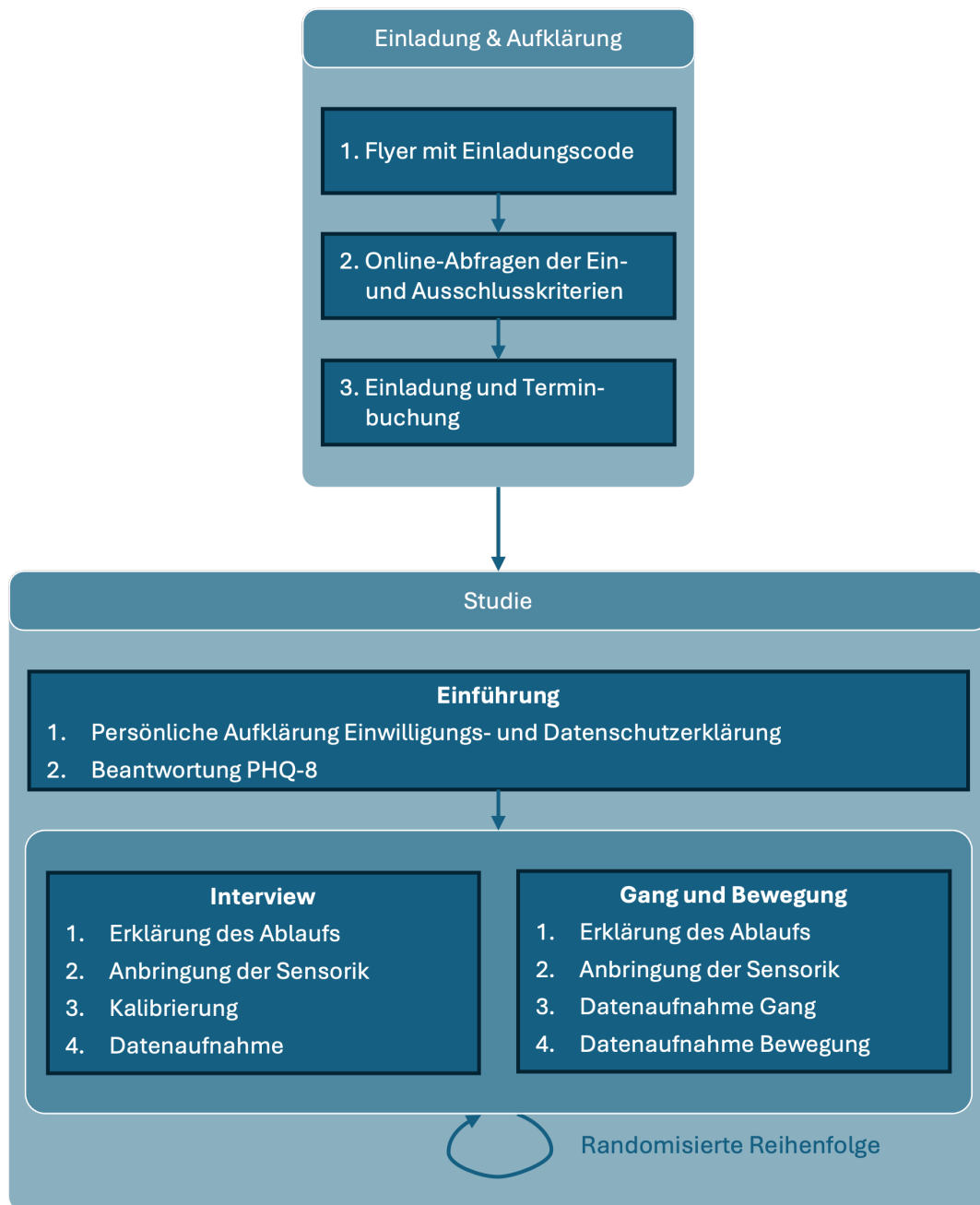


Abbildung 1: Schematischer Ablauf der Studie

B) Aufnahme der Gang- und Bewegungsdaten

- Individuelle Anpassung der Hörerlänge und Hörgerät-Schirmchens (Dome)
- Anbringen der Sensorik (Hörgerät) am Ohr (nur rechts)
- Das Hörgerät hat keinerlei Verstärkung und beeinflusst/verändert das Hörvermögen der Proband:innen nicht.
- Das Hörgerät wird über Bluetooth mit einem Smartphone und der dazugehörigen App gekoppelt, um Sensordaten aufzunehmen.

- Gang-Aufnahme: Proband:in geht auf einem festgesetzten Pfad mit selbstgewählter Geschwindigkeit.
- Anschließend werden allgemeine Kopfbewegungen (circa 2-3 Minuten) aufgenommen. Dafür wird der/die Proband:in in ein Gespräch über den Verlauf der bisherige Studie verwickelt.

Während eines Termins führen die Proband:innen die Aufgaben mit den angebrachten Sensoren unter Laborbedingungen aus. Im ersten Teil werden für die Tonaufnahme Hörgeräte-Dummies verwendet, die per Kabel mit einem Aufnahmegerät auf einem Tisch verbunden sind, sowie ein Stirnband mit Sensoren, die kabellos mit einem Mobiltelefon zur Aufzeichnung von Kopfbewegungen verbunden sind. Zunächst werden einfache Bewegungen zur Kalibrierung der Bewegungsachsen aufgezeichnet. Danach beginnt die Aufnahme der Audiodaten, bei dem die Teilnehmer:innen verschiedene Fragen in Form eines Interviews beantworten.

Im zweiten Teil werden die Testpersonen dazu aufgefordert, sich auf einem vorgegebenen Pfad eine bestimmte Zeit zu bewegen. Die Art der Fortbewegung (z. B. Geschwindigkeit, Gang) steht ihnen dabei frei. Zuletzt werden den Proband:innen Fragen gestellt, während Daten zur Bewegung des Kopfes aufgezeichnet werden.

Es werden bei den Teilnehmenden anthropometrische und demografische Daten erfasst. Zusätzlich wird ein standardisierter Fragebogen zur Erfassung des Vorliegens einer depressiven Symptomatik (Patient Health Questionnaire - PHQ-8 [12]) durchgeführt.

7.4. Zeitlicher Ablauf (Termine) und Studiendauer für den/die einzelne:n Proband:in
Das Experiment dauert ungefähr 60 – 75 Minuten:

- ca. 10 Minuten zur allgemeinen Vorbereitung und zum Ausfüllen des Fragebogens
- ca. 5 Minuten zur Anbringung und Kalibration der Sensoren
- ca. 20 Minuten zur Durchführung des Interviews
- ca. 5 Minuten zum Wechsel der Sensoren
- ca. 15 Minuten zur Aufnahme der Gang- und Bewegungsdaten

7.5. Gesamtdauer der Studie

Die Studie findet von Februar 2025 bis November 2025 statt.

8. Nutzen-Risiko-Abwägung (Sicherheitsaspekte)

8.1. Studienbedingte Risiken

Es bestehen keine gesundheitlichen Risiken für die Teilnehmenden. Durch den zusätzlichen standardisierten Fragebogen (PHQ-8) entstehen keinerlei Risiken. Die körperlichen Belastungen inklusiv der Bewegungsanalyse entsprechen der Alltagsbelastungen oder normalen Belastungen, die von gesunden Personen ohne Probleme geleistet werden können.

Die Belastung durch die Untersuchung mit Beschleunigungssensoren oder anderen Biosensoren ist als minimal zu verstehen. Die Befestigung der Sensoren erfolgt so am Körper der Proband:innen, dass diese durch die Sensoren so wenig wie möglich in Ihrer normalen Verhaltensweise beeinträchtigt werden. Die Sensoren sind passiv, in sich abgeschlossen und mit nur sehr geringen Strommengen versorgt (Batterie & Akku), so dass hier keine Gesundheitsgefährdung zu erkennen ist. Eine Gefahr, ausgehend von den Sensoren, ist daher auszuschließen.

Da die Hörgeräte ohne Verstärkung genutzt werden, ist die Nutzung dieser für die Teilnehmenden nicht als Belastung zu sehen.

8.2. Mit der Studie verbundener Nutzen

Für die Studienteilnehmer

Es ergibt sich kein direkter Nutzen für die Studienteilnehmenden. Die Ergebnisse der Studie können zu einem zukünftigen Nutzen für die Zielgruppe beitragen durch die Erweiterung der normalen Funktionalität von Hörgeräten mit Sprach- und Bewegungsanalysen.

Für die Wissenschaft

Die entwickelten Algorithmen zur Bewertung von depressiven Symptomen können in die Wissenschaft für unterschiedliche Zwecke implementiert werden: Diagnose, Früherkennung von Depression im häuslichen Umfeld, zur Verbesserung der Gesundheit, und zur Weiterentwicklung von Interventionen.

8.3. Abbruchkriterien

Ein Studienabbruch für den/die individuelle/n Probanden:in erfolgt, wenn:

- ein Ereignis eintritt, welches zur Gefährdung des/der Probanden:in oder des Personals führen kann,
- der/die Proband:in das Einverständnis zur Teilnahme an der Studie zurückzieht,
- der/die Proband:in sich nicht an die Anweisungen der Studiendurchführer:innen hält.

Eine Studie kann durch den/die Proband:in jederzeit und ohne Angabe von Gründen abgebrochen werden.

8.4. Statement zur ärztlichen Vertretbarkeit

Die Risiken sind, wie oben beschrieben, sehr gering. Durch die Studie werden Algorithmen zur Auswertung von Biosignalen entwickelt. Somit ist zur Durchführung der Studienmaßnahmen keine ärztliche Betreuung notwendig.

9. Datenmanagement und Datenschutz

9.1. Datenerfassung, -speicherung

Alle relevanten Dokumente der Studie werden aufgehoben, einschließlich (aber nicht beschränkt auf) Protokoll, der Einverständnis- und Datenschutzerklärung der Proband:innen und des Schriftwechsels mit der Ethik-Kommission, und zwar für eine Periode von zehn (10)

Jahren, folgend dem Datum des Endes der Studie. Alle Daten werden pseudonymisiert basierend auf einer Proband:innen-ID und werden erst dann auf einem elektronischen Medium gespeichert und zur weiteren Auswertung verwendet.

9.2. Anonymisiert/pseudonymisiert

Die Namen der Proband:innen und alle anderen vertraulichen Informationen unterliegen der Schweigepflicht und den Bestimmungen des Bundesdatenschutzgesetzes (BDSG). Die wissenschaftliche Auswertung der Daten erfolgt in pseudonymisierter Form.

Der Verschlüsselungscode besteht aus einer 4-stelligen zufällig generierten Kombination aus Zahlen und Buchstaben. Die Proband:innen werden durch eine Kodierungsliste identifiziert, die nicht mit den persönlichen Daten verknüpft werden kann.

9.3. Datenweitergabe

Eine Datenweitergabe erfolgt an die beteiligte Einrichtung der FAU wie in der Einverständnis- und Datenschutzerklärung angegeben. Notwendige Daten werden nur auf Anfrage und in pseudonymisierter Form an die genannten Einrichtungen weitergegeben. Namen und alle anderen vertraulichen Informationen unterliegen der Schweigepflicht und der Bestimmung des Bundesdatenschutzgesetzes (BDSG). Die wissenschaftliche Auswertung der Daten erfolgt in pseudonymisierter Form. Eine Weitergabe an Dritte erfolgt nur unter der ausdrücklichen schriftlichen Zustimmung der Proband:innen. Es werden keine aufgenommenen Daten (Sprach- oder Bewegungsdaten) an den Kooperationspartner übermittelt. Die Ergebnisse der Auswertungen und Analysen werden mit dem Kooperationspartner geteilt.

9.4. Widerruf und Datenlöschung

Die Proband:innen werden darüber informiert, dass die Teilnahme an dieser Studie freiwillig ist und dass sie ihr Einverständnis jederzeit, ohne Angabe von Gründen und ohne Nachteile für ihre weitere medizinische Versorgung zurückziehen können. Im Fall eines solchen Widerrufs erklären sich die Proband:innen einverstanden, dass die bis dahin aufgenommenen Daten ohne Namensnennung weiterhin verwendet werden dürfen, soweit dies erforderlich ist.

10. Umgang mit Biomaterialien

Es erfolgt keine Entnahme von Biomaterialien.

11. Referenzen

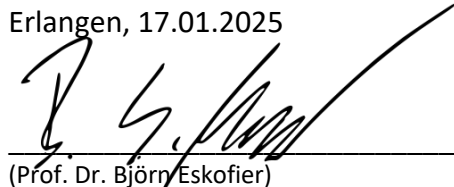
- [1] K. Kroenke, R. L. Spitzer, J. B. W. Williams, and B. Löwe, "The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review," *Gen. Hosp. Psychiatry*, vol. 32, no. 4, pp. 345–359, Jul. 2010, doi: 10.1016/j.genhosppsy.2010.03.006.
- [2] "Depressive disorder (depression)." Accessed: Sep. 04, 2024. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/depression>
- [3] J. D. Bukh, C. Bock, M. Vinberg, and L. V. Kessing, "The effect of prolonged duration of untreated depression on antidepressant treatment outcome," *J. Affect. Disord.*, vol. 145, no. 1, pp. 42–48, Feb. 2013, doi: 10.1016/j.jad.2012.07.008.
- [4] D. C. Mohr, M. Zhang, and S. M. Schueller, "Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning," *Annu. Rev. Clin. Psychol.*, vol. 13, no. 1, pp. 23–47, May 2017, doi: 10.1146/annurev-clinpsy-032816-044949.
- [5] P. Wu, R. Wang, H. Lin, F. Zhang, J. Tu, and M. Sun, "Automatic depression recognition by intelligent speech signal processing: A systematic survey," *CAAI Trans. Intell. Technol.*, vol. 8, no. 3, pp. 701–711, 2023, doi: 10.1049/cit2.12113.
- [6] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investig. Otolaryngol.*, vol. 5, no. 1, pp. 96–116, 2020, doi: 10.1002/lio2.354.
- [7] H. Hollien, "Vocal Indicators of Psychological Stress," *Ann. N. Y. Acad. Sci.*, vol. 347, no. 1, pp. 47–72, 1980, doi: 10.1111/j.1749-6632.1980.tb21255.x.
- [8] R. Feldman, S. Schreiber, C. Pick, and E. Been, "Gait, Balance and Posture in Major Mental Illnesses: Depression, Anxiety and Schizophrenia," Jan. 2020.
- [9] E. Garcia-Ceja, M. Riegler, T. Nordgreen, P. Jakobsen, K. J. Oedegaard, and J. Tørresen, "Mental health monitoring with multimodal sensing and machine learning: A survey," *Pervasive Mob. Comput.*, vol. 51, pp. 1–26, Dec. 2018, doi: 10.1016/j.pmcj.2018.09.003.
- [10] A.-K. Seifer, E. Dorschky, A. Küderle, H. Moradi, R. Hannemann, and B. M. Eskofier, "EarGait: Estimation of Temporal Gait Parameters from Hearing Aid Integrated Inertial Sensors," *Sensors*, vol. 23, no. 14, Art. no. 14, Jan. 2023, doi: 10.3390/s23146565.
- [11] T. Röddiger *et al.*, "Sensing with Earables: A Systematic Literature Review and Taxonomy of Phenomena," *Proc ACM Interact Mob Wearable Ubiquitous Technol*, vol. 6, no. 3, p. 135:1-135:57, Sep. 2022, doi: 10.1145/3550314.
- [12] K. Gräfe, S. Zipfel, W. Herzog, and B. Löwe, "Screening psychischer Störungen mit dem "Gesundheitsfragebogen für Patienten (PHQ-D)"", *Diagnostica*, vol. 50, no. 4, pp. 171–181, Oct. 2004, doi: 10.1026/0012-1924.50.4.171.
- [13] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *J. Affect. Disord.*, vol. 114, no. 1, pp. 163–173, Apr. 2009, doi: 10.1016/j.jad.2008.06.026.
- [14] W. Kong, S. Sessa, M. Zecca, and A. Takanishi, "Anatomical Calibration through Post-Processing of Standard Motion Tests Data," *Sensors*, vol. 16, no. 12, Art. no. 12, Dec. 2016, doi: 10.3390/s16122011.

12. Unterschriften: Studienleiter:in (Antragsteller:in)

Ich weiß, dass auch bei einer positiven Beurteilung des Vorhabens durch die Ethik-Kommission der Medizinischen Fakultät der FAU Erlangen-Nürnberg die wissenschaftliche ärztliche und juristische Verantwortung für die Durchführung des Projektes uneingeschränkt bei der Leiterin/dem Leiter verbleibt.

Unterschrift der Antragsteller:innen und Studienleiter:innen

Erlangen, 17.01.2025



(Prof. Dr. Björn Eskofier)

Lehrstuhl für Maschinelles Lernen und Datenanalytik

(Ann-Kristin Seifer)

Lehrstuhl für Maschinelles Lernen und Datenanalytik

Appendix B

Study Procedure

Teil A: Interview

Das Interview soll als Gespräch auf Alltagsebene stattfinden. Dabei sollen generell „positive“ und „negative“ Situationen abgefragt werden, ohne dabei vielleicht zu tiefgreifend zu werden. Die folgenden Fragen können als Anhaltspunkte verwendet werden, auf die jeweiligen Antworten der Teilnehmenden soll aber auch individuell eingegangen werden.

Vorbereitung

- Hörgerät einschalten und mit der App koppeln
- Stirnband an Proband:in anlegen und Hörgerät an der Seite fixieren
- Fotos der HA-Positionen machen
- Aufnahme per App starten
- Kalibrierung der Achsen durch folgende Aufgaben
 1. [Label *Sitting* setzen] 20 Sekunden aufrecht stehen (Person auffordern, 20 Sekunden einfach zu stehen) [Label *Nothing* setzen]
 2. [Label *Sitting* setzen] 3x *Sit-To-Stand*-Test (Person auffordern, sich dreimal hintereinander aufzustehen und hinzusetzen) [Label *Nothing* setzen]
- Dummy-Hörgeräte an Ohren anbringen: „Ich werde ein paar Fragen stellen und Sie dürfen darauf einfach ehrlich antworten. Bei der Studie interessiert uns nicht, was gesagt wird. Wir benötigen eher möglichst viele Daten zu verschiedenen Stimmmerkmalen wie Tonhöhe etc., deshalb wäre es schön, wenn Sie möglichst viel und ausführlich reden.“
- Aufnahme beim Recorder betätigen

Durchführung

Fragenkatalog		✓
Schön, dass Sie bei unserer Studie teilnehmen. Erzählen Sie uns gerne ein bisschen, wie es Ihnen heute geht. Warum fühlen Sie sich so?		
	Wie geht es Ihnen damit? Wie kommen Sie mit diesem Gefühl zurecht, wie wollen Sie mit diesem Gefühl umgehen?	
Wie ist es Ihnen in den letzten 2 Wochen ergangen? Wie erging es Ihnen körperlich und psychisch? Hatten Sie Beschwerden ?		
	Körperlich? Sport?	
	Psychisch?	
	Weitere Nachfragen zu Situationen stellen: Stress, Arbeit, Freizeit, ...	

Können Sie sich an gute oder schlechte Nachrichten erinnern, die Sie in den letzten Wochen erhalten haben?		
	Wie haben Sie sich dabei gefühlt?	
	Weitere Nachfragen zu Situationen stellen: Arbeit, Freizeit, ...	
Negative Gedanken formulieren		
	Gibt es Dinge, die Sie in letzter Zeit traurig gemacht haben?	
	Was sind Dinge, die Sie wirklich wütend machen?	
	Können Sie spontan einen negativen Gedanken formulieren?	
Positive Gedanken formulieren		
	Was würden Sie sagen, sind einige Ihrer besten Eigenschaften? Wie würden Freunde Sie beschreiben?	
	Gibt es Dinge, die Sie in letzter Zeit positiv gestimmt haben?	
	Was sind die Dinge, die Sie normalerweise in gute Stimmung versetzen?	
	Können Sie spontan einen positiven Gedanken formulieren?	
Erzählen Sie uns eine Geschichte aus Ihrer Kindheit/Jugend		
	Gibt es eine unangenehme Erinnerung, die Sie haben?	
	Was ist ein schöner Tag, an den Sie sich erinnern (z. B. bestimmtes Weihnachtsfest, Geburtstag, Urlaub, ...)?	
	Können Sie sich an das Gefühl erinnern, als Sie volljährig wurden? War da ein Gefühl der Freiheit oder eher Druck?	
Bild beschreiben		
	Beschreiben Sie das Bild (siehe Anhang). Was könnte die Geschichte dahinter sein?	
		
Text vorlesen (Max Frisch: Homo faber. Ein Bericht)		
	Lesen Sie den Text laut vor (siehe Anhang).	
Wie sieht in Ihrer kommenden Woche an?		
	Gibt es etwas, auf das Sie sich freuen?	
Zum Abschluss nochmal die Frage: Wie geht es Ihnen gerade? Hat sich während des Gesprächs etwas in Ihrer Gefühlslage geändert?		

Vielen Dank für das Gespräch. Somit wäre der erste Teil unserer Studie geschafft und wir kommen zum nächsten Teil.	
--	--

Teil B1: Gangdatenaufnahme

- Hörgeräte eventuell anpassen
- Hörgeräte mit Batterie und einschalten
- Hörgeräte mit App verbinden
- Verstärkung überprüfen, ob aus
- Sagen: „Das Tragen ist im ersten Moment vielleicht ungewohnt, aber man gewöhnt sich daran. Die Verstärkung ist zu jederzeit ausgeschaltet.“
- Hörgeräte anlegen
- Fotos der HA-Positionen und Person mit Kleidung machen
- [Label *Walking* setzen] Handy übergeben: „Bitte behalten Sie es bei sich.“
- Laufen Sie 3 Runden um das Gebäude, kommen Sie dann wieder zurück

Teil B2: Bewegungsdatenaufnahme

- Handy abnehmen, Sensoren unbedingt noch tragen: „Ich nehme schon mal das Handy, die Sensoren bitte noch kurz anlassen“ [Label *Nothing* setzen]
- [Label *Talking* setzen] „Anbei noch ein Fragebogen für Feedback, den Sie anonymisiert ausfüllen dürfen. Ich gebe Ihnen ein paar Minuten Zeit zum Ausfüllen und komme dann wieder zurück.“ (siehe Anhang)

	1	2	3	4	5
Wahrnehmung der Aufgabe					
Die bisherige Aufgabe war verständlich.					
Ich habe mich während der bisherigen Aufgabe wohlfühlt.					
Ich hatte das Gefühl, dass ich mich natürlich und frei bewegen konnte.					
Reflexion über Verhalten					
Ich habe mich bemüht, meine Bewegungen besonders gleichmäßig zu halten.					
Mir war bewusst, wie ich mich während der Aufgabe bewegt habe.					

Ich habe versucht, mein Verhalten an die Aufgabe anzupassen.					
Allgemeiner Eindruck					
Die bisherigen Aufgaben haben mir gefallen.					
Die Aufgaben waren nicht anstrengend für mich.					
Die Aufgaben waren interessant.					

Beobachtungsfragen	✓
Haben Sie während der Aufgabe etwas bemerkt, das Ihnen besonders aufgefallen ist?	
Haben Sie etwas an Ihrer Bewegung oder Ihrem Verhalten verändert?	

- [Label *Nothing* setzen] Hörgeräte abnehmen
- Aufnahme beenden

Sonstiges

Wetterlage / Temperatur / Bodenbeschaffenheit / Außergewöhnliches

Max Frisch

Homo faber

Ein Bericht

Wir starteten in La Guardia, New York, mit dreistündiger Verspätung infolge Schneestürmen. Unsere Maschine war, wie üblich auf dieser Strecke, eine Super-Constellation. Ich richtete mich sofort zum Schlafen, es war Nacht. Wir warteten noch weitere vierzig Minuten draußen auf der Piste, Schnee vor den Scheinwerfern, Pulverschnee, Wirbel über der Piste, und was mich nervös machte, so daß ich nicht sogleich schlief, war nicht die Zeitung, die unsere Stewardess verteilte, *First Pictures Of World's Greatest Air Crash In Nevada*, eine Neuigkeit, die ich schon am Mittag gelesen hatte, sondern einzig und allein diese Vibration in der stehenden Maschine mit laufenden Motoren - dazu der junge Deutsche neben mir, der mir sogleich auffiel, ich weiß nicht wieso, er fiel auf, wenn er den Mantel auszog, wenn er sich setzte und sich die Bügelfalten zog, wenn er überhaupt nichts tat, sondern auf den Start wartete wie wir alle und einfach im Sessel saß, ein Blonder mit rosiger Haut, der sich sofort vorstellte, noch bevor man die Gürtel geschnallt hatte. Seinen Namen hatte ich überhört, die Motoren dröhnten, einer nach dem andern auf Vollgasprobe –

Ich war todmüde.

Ivy hatte drei Stunden lang, während wir auf die verspätete Maschine warteten, auf mich eingeschwatzt, obschon sie wußte, daß ich grundsätzlich nicht heirate.

Ich war froh, allein zu sein.

Endlich ging's los –

Ich habe einen Start bei solchem Schneetreiben noch nie erlebt, kaum hatte sich unser Fahrgestell von der weißen Piste gehoben, war von den gelben Bodenlichtern nichts mehr zu sehen, kein Schimmer, später nicht einmal ein Schimmer von Manhattan, so schneite es. Ich sah nur das grüne Blinklicht an unsrer Tragfläche, die heftig schwankte, zeitweise wippte; für Sekunden verschwand sogar dieses grüne Blinklicht im Nebel, man kam sich wie ein Blinder vor.

Rauchen gestattet.

Er kam aus Düsseldorf, mein Nachbar, und so jung war er auch wieder nicht, anfangs Dreißig, immerhin jünger als ich; er reiste, wie er mich sofort unterrichtete, nach Guatemala, geschäftlich, soviel ich verstand –

Wir hatten ziemliche Böen.



Feedback-Fragebogen zur Studie

Beantworten Sie die folgenden Fragen auf einer Skala von 1 – 5, dabei steht 1 für „stimme überhaupt nicht zu“ und 5 für „stimme vollkommen zu“.

	1	2	3	4	5
Wahrnehmung der Aufgabe					
Die bisherige Aufgabe war verständlich.					
Ich habe mich während der bisherigen Aufgabe wohlfühlt.					
Ich hatte das Gefühl, dass ich mich natürlich und frei bewegen konnte.					
Reflexion über Verhalten					
Ich habe mich bemüht, meine Bewegungen besonders gleichmäßig zu halten.					
Mir war bewusst, wie ich mich während der Aufgabe bewegt habe.					
Ich habe versucht, mein Verhalten an die Aufgabe anzupassen.					
Allgemeiner Eindruck					
Die bisherigen Aufgaben haben mir gefallen.					
Die Aufgaben waren nicht anstrengend für mich.					
Die Aufgaben waren interessant.					

Folgende Fragen sind zur offenen Beantwortung.

Haben Sie während der Aufgabe etwas bemerkt, das Ihnen besonders aufgefallen ist?
Haben Sie etwas an Ihrer Bewegung oder Ihrem Verhalten verändert?

Gesundheitsfragebogen

Patientenkennung: _____

Wie oft fühlten Sie sich im Verlauf der **letzten 2 Wochen** durch die folgenden Beschwerden beeinträchtigt?

	Überhaupt nicht	An einzelnen Tagen	An mehr als der Hälfte der Tage	Beinahe jeden Tag
Wenig Interesse oder Freude an Ihren Tätigkeiten				
Niedergeschlagenheit, Schwermut oder Hoffnungslosigkeit				
Schwierigkeiten ein- oder durchzuschlafen oder vermehrter Schlaf				
Müdigkeit oder Gefühl, keine Energie zu haben				
Verminderter Appetit oder übermäßiges Bedürfnis zu essen				
Schlechte Meinung von sich selbst; Gefühl, ein Versager zu sein oder die Familie enttäuscht zu haben				
Schwierigkeiten, sich auf etwas zu konzentrieren, z.B. beim Zeitungslesen oder Fernsehen				
Waren Ihre Bewegungen oder Ihre Sprache so verlangsamt, dass es auch anderen auffallen würde? Oder waren Sie im Gegenteil „zappelig“ oder ruhelos und hatten dadurch einen stärkeren Bewegungsdrang als sonst?				

Appendix C

Hyperparameter Search Space

Type	Algorithm	Parameter	Values
Feature Selection	SkB	k	{2, 4, ..., 18, "all"}
	RFE	n_features_to_select	{2, 4, ..., 18, None}
Classifier	LR	penalty	["l1", "l2"]
		C	{0.1, 1.0, 10.0}
	kNN	n_neighbors	{2, 3, 4}
	RF	n_estimators	{20, 40, ..., 300}
		criterion	["entropy", "gini"]
		max_features	{0.1, 0.2, 0.3, 0.4, "sqrt"}
		min_impurity_decrease	{0, 0.01, ..., 0.09}
		max_depth	{2, 4, ..., 20}
	SVM-lin	C	{ 10^{-2} , 10^{-1} , ..., 10^2 }
	SVM-rbf	C	{ 10^{-2} , 10^{-1} , ..., 10^2 }
		gamma	{ 10^{-2} , 10^{-1} , ..., 10^2 }
	XGBoost	n_estimators	{20, 30, ..., 100}
		max_depth	{1, 2, 3}
		learning_rate	{0.1, 0.2, ..., 1.0}
		reg_lambda	{0.1, 1.0, 10.0}

Table C.1: Grid-search parameter values for feature selection and classification algorithms.

Appendix D

Acronyms

AUC area under the curve

AVEC Audio/Visual Emotion Challenge

BDI-II Beck Depression Inventory-II

CNN convolutional neural network

CV cross-validation

DAIC Distress Analysis Interview Corpus

DL deep learning

DNN deep neural network

EEG electroencephalogram

FAU Friedrich-Alexander-Universität Erlangen-Nürnberg

GeMAPS Geneva Minimalistic Acoustic Parameter Set

GMM gaussian mixture model

HA hearing aid

HAM-D Hamilton Rating Scale for Depression

HNR harmonics-to-noise ratio

IMU	inertial measurement unit
kNN	k-nearest neighbors
LLD	low level descriptors
LR	logistic regression
LSTM	long short-term memory
ML	machine learning
MLP	multi-layer perceptron
MFCC	mel frequency cepstral coefficients
NHR	noise-to-harmonics ratio
PCA	principal component analysis
PHQ-8	Patient Health Questionnaire-8
PPG	photoplethysmogram
ResNet	residual neural network
RF	random forest
RIC	receiver in the canal
RMSE	root mean square error
RNN	recurrent neural network
SDR	speech depression recognition
SHAP	SHapley Additive exPlanations
SVM	support vector machine
XGBoost	extreme gradient boosting