



University of St. Gallen

School of Computer Science

to obtain the title of

Master of Science in Computer Science

Master Thesis on

Audio-centered Approach for Building a Multimodal Predictive AI Agent to Detect Depressive Behaviors

Submitted by:

Jonas Länzlinger

23-606-650

Approved on the application by:

Prof. Dr. Bruno Rodrigues

Date of Submission:

June 26, 2025

Acknowledgments

I express my gratitude to my master's thesis advisor, Prof. Dr. Bruno Rodrigues, for his guidance, constructive criticism, and encouragement. I am thankful for the freedom he gave me to explore directions that attracted my interests and for his consistent support along this journey, which has contributed to both the academic quality of this final work and my personal growth.

Furthermore, I want to thank Prof. Dr. Simon Mayer for his cosupervision on this master's thesis.

I extend my acknowledgments to the creators of the DAIC-WOZ dataset for allowing me to use their valuable scientific resources to conduct my research and evaluation endeavors on automated depression detection presented in this work.

Lastly and most importantly, my gratitude goes to my partner. Your patience and belief in me made all the difference.

St. Gallen, June 26, 2025

Jonas Länzlinger

Abstract

Depression remains one of the most serious and widespread mood disorders, affecting more than 280 million people worldwide. Current diagnostic approaches rely heavily on subjective self-reports and clinical interviews that are still failing to reliably capture the true patient's behavior in natural, everyday settings.

This master's thesis introduces a novel approach for automated mental health monitoring. Particularly designed around an acoustic-based approach for depression detection, designed specifically as a software application for IoT-enabled private households. Using passive sensing techniques, the system focuses on the detection of potential depressive behavior to allow timely intervention. By constructing a direct mapping between behavioral patterns and observable clinical symptoms, users can gain insight into their mental health state, helping to overcome the limitations of traditional methods.

The first contribution is the design of a conceptualized *System Architecture* for a modular, scalable system capable of automated mental health monitoring. The architecture aims to simplify the faced challenges involved in setting up such systems. Given its exploratory nature, the *System Architecture* pays attention in particular to the integration and flexible combination of multimodal data streams. It also guides temporal context modeling of personalized user behavior analysis that accounts for individual variability.

The second contribution is a theoretical *Linkage Framework* that unifies the linkage (mapping) of low-level feature data, such as acoustic voice features, to higher-level clinically grounded symptoms for mental health disorders. This proposes a unified, structured approach to bridge the gap between medical psychology and computer science, enabling automated reasoning about mental health states. This framework gets instantiated as a concrete *MDD Mapping* that links acoustic voice features to clinically interpretable indicators of Major Depressive Disorder (MDD), as defined by DSM-5.

The third contribution is the combination of contributions one and two: The concrete *MDD Mapping* gets employed within the instantiation of the proposed *System Architecture* as a prototypical software implementation.

Finally, the solution is assessed by evaluating (1) the functional and qualitative maturity of the *System Architecture* and (2) the practical soundness of the concrete *MDD Mapping*.

Table of Contents

Abstract	ii
List of Abbreviations	vi
List of Algorithms	x
List of Figures	xi
List of Tables	xiv
1 Introduction	1
1.1 Motivation	1
1.2 Problem Statement	3
1.3 Research Question	7
1.4 Methodology	9
2 Fundamentals	11
2.1 Background	11
2.1.1 Emotions	11
2.1.2 Classification Systems	12
2.1.3 Depression	13
2.1.4 Medical Assessment	14
2.1.5 Tone-based Approach	19
2.2 Related Work	21
2.2.1 ML-based Depression Detection	21
2.2.2 Modality	22
2.2.2.1 Text	22
2.2.2.2 Audio	24
2.2.2.3 Video	27
2.2.3 Multimodality	28
3 Design	30
3.1 Structured Approach	30

3.2	Linkage Framework	31
3.2.1	Conceptual Idea	31
3.2.2	Mapping Pyramid	32
3.3	MDD Mapping	33
3.4	Feature Selection	34
3.4.1	Depressed Mood	35
3.4.2	Loss of Interest	37
3.4.3	Significant Weight Changes	39
3.4.4	Insomnia / Hypersomnia	41
3.4.5	Psychomotor Retardation / Agitation	43
3.4.6	Fatigue / Loss of Energy	44
3.4.7	Worthlessness / Guilt	46
3.4.8	Diminished Ability to Think / Concentrate	48
3.4.9	Suicidality	50
3.5	Feature Mapping Specification	52
3.5.1	Feature Exploration	54
3.5.2	Audio Descriptors	55
3.5.3	Shared Information	58
3.5.4	Extraction	59
3.5.5	Analysis Approach	60
4	Implementation	61
4.1	Scenario Description	61
4.2	Requirement Engineering	62
4.2.1	User Stories	63
4.2.2	Functional Requirements	64
4.2.3	Non-functional Requirements	67
4.3	Architecture	68
4.3.1	Architectural Style	68
4.3.2	Design Instantiation	70
4.3.2.1	High-level Design	70
4.3.2.2	Data Ingestion Layer	74
4.3.2.3	Bridge Layer	80
4.3.2.4	Processing Layer	81

4.3.2.5	Persistence Layer	84
4.3.2.6	Temporal Context Modeling Layer	85
4.3.2.7	Analysis Layer	91
4.3.2.8	Dashboard Layer	101
5	Evaluation	104
5.1	System Architecture	104
5.1.1	Functional Assessment	104
5.1.2	Non-functional Assessment	108
5.2	Linkage Framework	116
5.2.1	Method	117
5.2.2	Hypotheses	117
5.2.3	DAIC-WOZ Dataset	119
5.2.4	Hypotheses Testing	120
5.2.5	Second Attempt	125
5.3	Discussion	127
6	Considerations & Future Work	130
6.1	Summary	130
6.2	Considerations	131
6.3	Future Work	133
	Bibliography	135
	Appendix A Questionnaires	154
	Appendix B Prototype	158
	Appendix C Evaluation	166
	Appendix D GitHub Repository	170
	Declaration of Authorship	171
	Declaration of Auxiliary Aids	172

List of Abbreviations

AI	Artificial Intelligence
AM	Amplitude Modulation
APA	American Psychiatric Association
API	Application Programming Interface
AU12/14/15	Facial Action Units
AUC-ROC	Area Under the (Curve) – Receiver-operating Characteristic Curve
BDNF	Brain-derived Neurotrophic Factor
BERT	Bidirectional Encoder Representations from Transformers
BF	Body Fat
BMI	Body Mass Index
CBT	Cognitive Behavioral Therapy
CI/CD	Continuous Integration / Continuous Delivery
CNN	Convolutional Neural Network
ComParE16	INTERSPEECH 2016 Computational Paralinguistics Challenge
COVID-19	Coronavirus Disease 2019
csHMM	Context-sensitive Hidden Markov Model
DAIC-WOZ	Distress Analysis Interview Corpus – Wizard of Oz
DNA	Deoxyribonucleic Acid
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
DSR	Design Science Research
DST	Dexamethasone Suppression Test

EMA	Exponential Moving Average
FCN	Fully Convolutional Network
FDR	False Discovery Rate
FER	Facial Emotion Recognition
FR	Functional Requirement
GeMAPS	Geneva Minimalistic Acoustic Parameter Set
GRBS	Guilt-related Brain Signature
EEG	Electroencephalogram
EMG	Electromyogram
EOG	Electrooculogram
HD	Hypothetico-Deductive
HLD	High-level Descriptor
HMM	Hidden Markov Model
HNR	Harmonic-to-Noise Ratio
HPA	Hypothalamic-pituitary-adrenal
HSAT	Home Sleep Apnea Test
ICD-11	International Classification of Diseases, 11th Revision
IoT	Internet of Things
LLD	Low-level Descriptor
LLM	Large Language Model
LSTM	Long Short-term Memory
LSVM	Linear Kernel Support Vector Machine
MCI	Mild Cognitive Impairment

MDD	Major Depressive Disorder
MDL	Multimodal Depressive Dictionary Learning
ML	Machine Learning
MQTT	Message Queuing Telemetry Transport (formerly)
MSA	Microservices Architecture
NFR	Non-functional Requirement
NLP	Natural Language Processing
NN	Neural Network
OSN	Online Social Network
PAT	Peripheral Arterial Tonometry
PDD	Psychological Disorder Detection
PDF	Probability Density Functions
PHQ-8	Patient Health Questionnaire – 8
PHQ-9	Patient Health Questionnaire – 9
PMA	Psychomotor Agitation
PMR	Psychomotor Retardation
PRISMA	Preferred Reporting Items for Systematic Review and Meta-Analysis
PSD	Power Spectral Density
PSG	Polysomnography
PTSD	Post-traumatic Stress Disorder
QoS	Quality of Service
RE	Requirement Engineering
REST	Representational State Transfer

RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SCID-5	Structured Clinical Interview for DSM-5
SDD	Speech Depression Detection
SDI	Standardized Diagnostic Interview
SER	Speech Emotion Recognition
SF	Spectral Flatness
SMN	Social Media Network
SNR	Signal-to-Noise Ratio
STFT	Short-time Fourier Transform
STM	Spectro-temporal Modulation
SVM	Support Vector Machine
TESS	Toronto Emotional Speech Set
TST	Total Sleep Time
US	User Story
VAD	Voice Activity Detection
VOT	Voice Onset Time
VSA	Vowel Space Area
WHO	World Health Organization
ZCR	Zero-crossing Rate

List of Algorithms

4.1	Smoothing factor for EMA [28].	88
4.2	EMA update computation [28].	88
4.3	Variation – spike-dampened EMA computation.	89
4.4	<i>Z-score</i> computation [220].	93
4.5	Direction modifier function.	97
4.6	Indicator score computation.	97
4.7	Baseline finetuning function.	100
5.1	Formulated real-time processing constraint.	111

List of Figures

1.1	Percentage of studies investigating text, audio, and video for automated depression detection [226].	5
1.2	Conceptual project idea.	8
2.1	Tri-dimensional emotion model by Latinjak [250].	12
2.2	Bimodal concept of knowledge in speech data [26].	20
2.3	Speech production process [162].	21
2.4	Voice feature effect sizes according to Low <i>et al.</i> [50].	25
3.1	<i>Linkage Framework, conceptual Mapping Pyramid.</i>	32
3.2	<i>Mapping Pyramid</i> instantiated with MDD in the <i>diagnosis layer.</i>	34
3.3	Mapping overlap between indicator (4) <i>Insomnia / hypersomnia</i> and (6) <i>Fatigue / loss of energy.</i>	35
3.4	Indicator (1) <i>Depressed mood</i> mapping.	36
3.5	Indicator (2) <i>Loss of interest</i> mapping.	38
3.6	Indicator (3) <i>Significant weight changes</i> mapping.	39
3.7	Indicator (4) <i>Insomnia / hypersomnia</i> mapping.	41
3.8	Indicator (5) <i>Psychomotor retardation / agitation</i> mapping.	44
3.9	Voice features between control and sleep deprived groups [58].	44
3.10	Indicator (6) <i>Fatigue / loss of energy</i> mapping.	45
3.11	STM projections between control (A) and sleep deprived (B) groups [58].	46
3.12	Indicator (7) <i>Worthlessness / guilt</i> mapping.	47
3.13	Indicator (8) <i>Diminished ability to think / concentrate</i> mapping.	49
3.14	Indicator (9) <i>Suicidality</i> mapping.	51
3.15	Voice metrics classification [27].	55
3.16	Semantic classification process for a <i>Violin</i> sound (HLD) [163].	56
3.17	<i>MDD Mapping</i> feature intersections – matrix plot.	58
3.18	<i>MDD Mapping</i> feature intersections – upset plot.	58
4.1	High-level architectural design – component diagram.	71
4.2	Modeled bounded contexts.	72
4.3	<i>AbstractEdgeDevice</i> – UML class diagram.	76
4.4	<i>Data ingestion layer</i> – UML class diagram.	77
4.5	Process logic of <i>data ingestion layer</i> – flow diagram.	79

4.6	Process logic of <i>bridge layer</i> – flow diagram.	80
4.7	Process logic of <i>processing layer</i> – flow diagram.	82
4.8	Implemented database collections model.	85
4.9	Process logic of <i>temporal context modeling layer</i> – flow diagram.	86
4.10	Process steps – (left) <i>aggregate metrics</i> and (right) <i>contextualize metrics</i>	90
4.11	Process logic of <i>analysis layer</i> , sub-process <i>metrics analysis</i> – flow diagram.	92
4.12	Z-score computation – explanatory visualization [220].	93
4.13	Process logic of <i>analysis layer</i> , sub-process <i>indicator derivation</i> – flow diagram.	95
4.14	Syntactic file structure of the <i>config.json</i> file [128].	96
4.15	Process logic of <i>analysis layer</i> , sub-process <i>baseline finetuning</i> – flow diagram.	98
4.16	Process steps – (left) <i>analyze metrics</i> and (right) <i>indicator derivation</i>	101
4.17	Process logic of <i>dashboard layer</i> – flow diagram.	102
4.18	<i>Dashboard</i> screenshots – (left) <i>raw metrics</i> and (right) <i>analyzed metrics</i>	103
4.19	<i>Dashboard</i> screenshots – (left) <i>indicator scores</i> and (right) <i>PHQ-9 questionnaire</i>	103
5.1	Illustration of horizontal scaling on the system’s <i>continuous side</i>	109
5.2	Performance test – processing duration measurements.	111
5.3	Baseline (<i>population_baseline.py</i>) specification file [128].	113
5.4	Simulated $F0_{avg}$ data – <i>triggered side</i> of data pipeline (1/2) [128].	114
5.5	Simulated $F0_{avg}$ data – <i>triggered side</i> of data pipeline (2/2) [128].	115
5.6	Conceptual idea of the <i>Linkage Framework</i>	116
5.7	Self-reported PHQ-8 indicator scores across participants.	121
5.8	Distribution of selected male and female acoustic voice features.	122
5.9	<i>Cohen’s d</i> (positive = higher in males, negative = higher in females).	123
5.10	Pearson correlation coefficients of feature-to-indicator pairings.	124
5.11	Pearson p-values of feature-to-indicator pairings.	125
5.12	DSM-5 <i>indicator scores</i> for simulated streamed data of TESS dataset.	126
A.1	PHQ-8 questionnaire (1/2) [110].	154
A.2	PHQ-8 questionnaire (2/2) [110].	155
A.3	PHQ-9 questionnaire (1/2) [186].	156
A.4	PHQ-9 questionnaire (2/2) [186].	157
B.1	Prototype screenshot (1/8) – Home [128].	158

B.2	Prototype screenshot (2/8) – Raw Metrics [128].	159
B.3	Prototype screenshot (3/8) – Aggregated Metrics [128].	160
B.4	Prototype screenshot (4/8) – Contextual Metrics [128].	161
B.5	Prototype screenshot (5/8) – Analyzed Metrics [128].	162
B.6	Prototype screenshot (6/8) – DSM-5 Indicator Scores [128].	163
B.7	Prototype screenshot (7/8) – Baseline [128].	164
B.8	Prototype screenshot (8/8) – PHQ-9 Questionnaire [128].	165
C.1	Acoustic voice feature sample distribution – scatter plot (1/3) [128]. . .	166
C.2	Acoustic voice feature sample distribution – scatter plot (2/3) [128]. . .	167
C.3	Acoustic voice feature sample distribution – scatter plot (3/3) [128]. . .	168
C.4	Adjusted Pearson p-values of feature-to-indicator pairings [128].	169
D.1	GitHub repository <i>depression-detection</i> [128].	170

List of Tables

1.1	Overview of research objectives.	9
2.2	DSM-5 coding for MDD [18].	14
2.1	DSM-5 diagnostic criteria for MDD [18].	15
2.3	PHQ-9 total score interpretation for depression severity [109].	17
2.4	Text-based depression detection studies.	24
2.5	Audio-based depression detection studies.	27
2.6	Video-based depression detection studies.	28
3.1	DSM-5 diagnostic criteria for the insomnia disorder [18].	41
3.2	Overview <i>Feature Mapping Specification</i> fields.	53
3.3	MDD <i>Feature Mapping Specification</i>	53
3.4	Voice features descriptor type classification.	57
4.1	Defined user stories.	63
4.2	Defined functional requirements.	64
4.3	Service-level architecture paradigms.	73
5.1	Checklist of implemented functional requirements.	105
5.2	Involved test files for the performance test [128].	110
5.3	Formulated hypotheses.	117
5.4	PHQ-8, PHQ-9 questionnaire items [109, 110] and the MDD DSM-5 clinical indicators [18].	120
5.5	Evaluation of directional hypotheses on acoustic voice features.	126
5.6	Overview of functional requirement (FR) evaluation.	127

Chapter 1

Introduction

1.1 Motivation

In today's world, mental health is at an alarming risk: The World Health Organization (WHO) reports that an estimated 3.8% of the global population (approximately 280 million people) suffer from various depressive disorders [175]. However, incidents are not evenly distributed between demographics: Women are affected approximately 50% more often than the male population; 10% of pregnant women who just gave birth experience episodes of depression [32]; 5.7% of older people over 60 years of age suffer from depression [175]. Although marital status and differences between rural and urban living environments have effects on the susceptibility to depression, there is no evidence of the prevalence of depression in any race or socioeconomic status [160]. Numerous studies demonstrate an alarming increase in depression incidents worldwide [15]. Today's higher rates of reported depression patients have ignited speculations about modern environments causing an 'epidemic of depression' [81].

Depression can begin early in life and often even be caused during the first years of a child's development. If the condition is not treated, it can introduce the risk of substance abuse, the development of chronic conditions, a decrease in life expectancy, or even premature death, such as suicide [174]. The numbers show that especially children are at risk: In the United States, almost 20% of children and adolescents between 3 and 17 years of age have a mental disorder, and suicidality among high school students has increased by more than 40% between the 2010s and 2020s; of course, COVID-19 played a key role in that regard [17]. In fact, self-harm and suicide are currently the fourth leading cause of death among people between 15 and 29 years of age [175]. Due to the critical stage of physical and mental development during childhood and adolescence, good mental health

and well-being are essential [17]. This is the reason why special efforts need to be made in research on depressive disorders, considering children and young adults in particular.

Often, depression is diagnosed together with one or more comorbidities such as substance use, panic disorder, and anxiety disorder [160]. Patients suffering from any form of depression are affected in many aspects of their life [175]. Firstly, people with depression characterize their subjective emotional state by an altered experience of emotions and body (negative emotions, lack of positive emotions, drained of energy, detached from the mind, the body, and the world), an altered experience of the self (purposelessness, existential hopelessness, losing control over one's thoughts, feeling numb, empty, dead), and an altered experience of time (*e.g.*, changes in vital biorhythms, stagnation of the present) [180]. Second, depression affects many domains of life, including relationships with friends and family [175] as well as lack of productivity at work [92]. However, depression affects not only those who suffer from it, but also those close to them. Individuals with a first-degree relative diagnosed with depression are three times more likely to develop the condition compared to the general population [38].

There are various approaches to the treatment of depression, such as cognitive behavioral therapy (CBT), pharmacotherapy, psychotherapy, interventional countermeasures, and lifestyle modification [159]. Unfortunately, effective treatments are largely accessible to people in high-income countries, where mental health services are better resourced and more widely available. In contrast, low- and middle-income countries often face significant challenges, including limited infrastructure, few trained professionals, or restricted access to care, leaving many patients without the appropriate support or treatment they might need [175]. Access barriers to psychotherapy can make it increasingly difficult for certain groups of people, (*e.g.*, low-income households or physically or mentally impaired people) to get their adequate help [174]. In fact, an interesting study by Lee *et al.* revealed several predictive factors for mental health care utilization, such as age, sex, and insurance coverage, and suggests reducing barriers for young adults, men, and uninsured / underinsured [240].

In contrast, the health care system is already under significant strain. A study by Schousboe *et al.* showed the impact of depressive disorders on a patient's total yearly health care costs. Individuals with depressive symptoms incur substantially higher annual costs – rising from \$4,654 for those with minimal symptoms to over \$9,000 for those with moderate to severe symptoms – even after adjustment for age, comorbidities, and functional limitations [101]. Furthermore, approximately a third of depression patients do

not achieve sustained relief even though they received multiple treatments, and especially patients who use disproportionately mental health care put pressure on general health care resources [5].

As access to mental health support remains largely limited to privileged groups and the strain on health care systems continuously grows, the critical question has to be asked: *"How do we find innovative and efficient methods to detect and manage depression more effectively during this 'epidemic of depression'?"* [15]

Society, researchers, and health care professionals are investigating substantial efforts to understand and address this question. For example, this situation is reflected in the significant increase in depression-related publications over the past decades [174].

Research provides evidence that early detection of mental disorders (such as depression) offers many benefits. Recognizing problems in the early stages allows the early application of an effective and suitable treatment, which has the potential for a better treatment outlook and long-term well-being, as well as an increase in cost-effectiveness [94, 155].

In this master's thesis, the fields of computer science and medical psychology are bridged to investigate this question. Advances in ubiquitous computing, driven by affordable hardware, flexible and powerful edge devices, the growing adoption of distributed architectures [131], and the emergence of tiny, lightweight, and efficient machine learning (ML) models [52], all build the technological foundation for the combination with proven medical evidence to make automated depression detection accessible to everyone, regardless of age, gender, socioeconomic status, cultural background, or geographic location.

1.2 Problem Statement

Although depression and suicidal cases are widespread, their diagnosis is difficult due to their complex nature and the wide variation of depression manifestations that are defined by individual sets of specific symptoms [162]. Therefore, the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) classifies a diverse range of different depressive disorders, each with its unique set of symptoms [18]. This master's thesis focuses on the detection of the most prevalent depressive disorder, Major Depressive Disorder (MDD).

Current diagnostic methods for assessing psychopathology still rely heavily on verbal reports such as clinical interviews, questionnaires, and self-reports. Although these

traditional assessment approaches are clinically useful, they also present some limitations and challenges [162]. Traditional depression assessment methods focus on the analysis of a set of specific categorical symptoms. Those symptoms are predominantly of a physical nature, which introduces a range of subjective biases when assigning scores to them [162]. Furthermore, the diagnosis of depression is a complex process influenced by many factors that originate both from the patient and the examiner [162]. This underlines the difficulty of accurately assessing depressive symptoms [42, 115]. Traditional diagnostic approaches lack systematic and efficient ways to integrate behavioral observations that may occur outside the patient's (or the examiner's) awareness [162]. Furthermore, self-reports, whether they concern past actions, hypothetical future behaviors, or perceived causes of behavior, have been shown to be unreliable indicators of actual behavior [190].

The limitations and challenges of the current state of depression diagnosis are problematic and may introduce biases or errors that could lead to a wrong diagnosis [162]. Currently, there is no objective measure for the detection of clinically useful depression, which poses a challenge to finding the optimal care and places a significant burden on the health, social, and economic system [162].

In response, a substantial amount of research has explored different approaches, traditional or not, for the diagnosis of depression [13, 76, 109, 115, 182]. One of the research directions is the field of automated depression detection that has the potential to improve the objectivity of depression assessments: it can assist clinicians in the decision-making process, identify early signs of depression, and track symptoms over time [112]. Automated depression detection is especially powerful because it offers the possibility to integrate multiple data modalities – such as text, audio, and video – each offering distinct and complementary insights that improve the accuracy and depth of the depression assessment [226]. Figure 1.1 illustrates the distribution of studies that focus on different data modalities, text, audio, and video, in the context of automated depression detection [226].

Many approaches to automated depression detection are based on statistics, ML, and artificial intelligence (AI) techniques [66, 248]. Those systems are often tailored for the analysis and investigation of a very specific but narrow set of features, thus lacking in modal variability [66, 248]. The survey by Yadav offers a structured review of research on automated depression detection, highlighting the scope and limitations of various approaches. The limitations are heavily dependent on the modality that has been used. For text analysis, language ambiguity, lack of context, and the inability to capture nonverbal

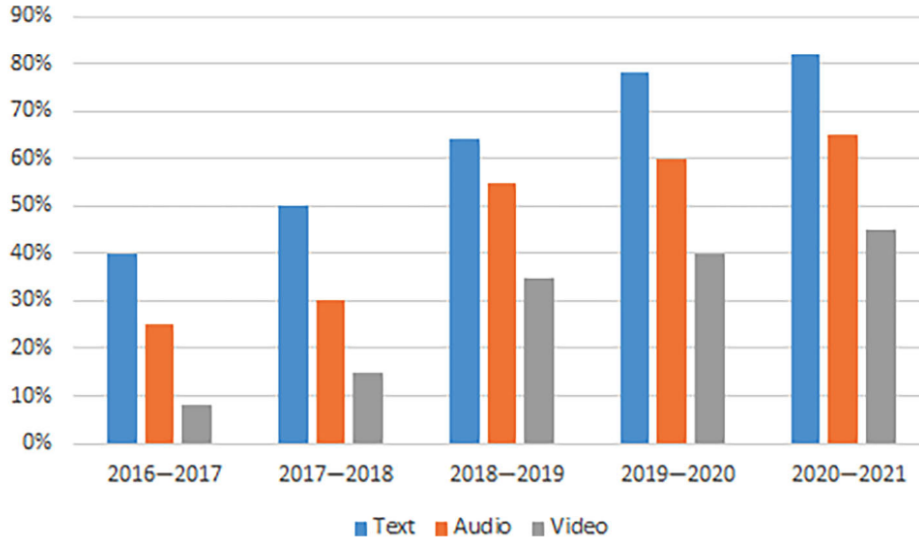


Figure 1.1: Percentage of studies investigating text, audio, and video for automated depression detection [226].

cues are the most prevalent limitations [226]. The audio detection is dominantly limited by scarce datasets with little variability, and most of the studies relied on the most basic audio features. And video depression detection studies often focused only on small sets of extracted features and underused modern powerful ML architectures [226].

Furthermore, these functional black-box machine learning approaches can successfully identify individuals at risk of depression but provide little to no insight into the decision-making process [66, 248]. Given that medical decisions greatly impact long-term treatment outcomes, it is essential to develop a transparent and explainable system for automated depression detection. Although such black boxes bring clinical value, the lack of interpretability limits trust in the system and provides no actionable insights for medical professionals.

One consistent insight across all modalities is that high-quality feature extraction is critical to the success of automated depression detection systems [226]. Consequently, research on identifying meaningful and discriminatory biomarkers of depressive behavior has recently gained significant popularity [162]. In other medical fields such as neurology, the usage of biomarkers is already well-established, while in psychiatry clinicians still rely heavily on traditional assessment methods without the integration of automated health assessment techniques, *e.g.*, through biomarkers [173].

Numerous biomarkers of various modalities have been investigated. Some biomarkers allow for reliable evaluation of behavioral and biological patterns associated with

psychiatric disorders, such as depression [161]. The focus of this master’s thesis is on the audio modality for automated depression detection. Current research focuses primarily on the linguistic biomarkers in this modality, where spoken text is analyzed. However, acoustic-based speech biomarkers have been shown to perform better than linguistic biomarkers in isolation [25]. One hypothesis might be that it is much easier to change *what* is said than *how* it is being said. Due to these scientific insights on the correlation between certain vocal biomarkers and depressive behavior [162], this master’s thesis revolves around an acoustic-based approach in an attempt to find promising acoustic-based voice features for automated depression detection.

However, it is also important to consider that people can still consciously or unconsciously modify their voice tone. For example, in a clinical diagnostic setting, they may alter the tone, speed, or rhythm in the presence of other people or under stress. This introduces undesirable biases (observer effects) that invalidate the authenticity of audio data [169]. For a reliable and accurate automated depression detection system, it is essential to capture ecologically valid speech samples that reflect the person’s natural vocal behavior.

To accomplish this, this work is contextualized within the scenario of a private household (as an IoT-enabled environment), which is assumed to be relatively free from external influences, allowing the collection of authentic speech data. Recording speech data in this setting reduces the risk that individuals intentionally modify their voice. This will minimize observer effects and holds the potential to yield authentic and better results [169]. Furthermore, this approach may reveal even stronger characteristics in acoustic voice biomarkers than in the case of structured interviews or laboratory studies [70].

However, there exists a gap in research that uses physiological data in real-world, daily life settings. The few existing studies that collected behavioral data in such contexts are often extremely time-consuming, making it difficult to automate these approaches or scale them effectively [190].

Yadav’s survey highlights another critical limitation in automated depression detection research: Most studies focus on a single modality and design feature extraction techniques and machine learning models specifically tailored to the data modality, context, and selected features of that modality, without considering integration with other modalities [226]. This narrow focus misses the opportunity to use multiple complementary data sources to improve prediction accuracy and reliability. This is where the proposed household setting is beneficial. The potential of an IoT-enabled environment opens the

possibility to incorporate multiple sensing data streams of varying modalities. Although this solution introduces a novel mechanism to the study of automated depression detection, it also faces exciting challenges from a software engineering perspective, particularly in terms of architectural design and the respective integration choices.

In summary, existing approaches to automated depression detection have made significant progress, but several challenges remain unresolved. This master's thesis introduces a novel automated depression detection approach, focusing on acoustic-based speech data. The system is designed specifically for IoT-enabled private household environments while prioritizing explainability, simplicity, and openness, laying the groundwork for a more holistic automated depression detection methodology. By allowing other researchers to easily extend the system with new modalities or features, this work promotes collaborative scientific innovation.

1.3 Research Question

The key research question explored in this master's thesis is formulated as follows:

How can an explainable white-box system for automated health assessments in IoT-enabled household environments be designed while promoting scientific collaboration and innovation; and to what extent can acoustic-based voice features – alone or in combination with other modalities – be integrated into that system to promote automated depression detection?

While the first part of the research question focuses on the conceptual and architectural design of an explainable system, the second part explores the role of acoustic-based voice features as potential biomarkers of depression. This overarching research question structures the master's thesis, with each chapter addressing specific objectives derived from the decomposition of that question. Each objective includes tasks that address given challenges to complete the given objective successfully. These objectives directly relate to the final contributions of this master's thesis that fill the scientific gap as described in Section 1.2. Figure 1.2 illustrates the conceptual idea showing how the individual objectives contribute to the overall scientific contribution of this master's thesis.

To answer the research question, an objective of this work is to develop a *Linkage Framework* that semantically maps mental health disorders to low-level biomarkers (*Obj-1*). This framework aims to ensure the interpretability of the system's decision-making process and to support the integration of multiple data modalities into a unified structure. Additionally, for *Obj-1*, this framework is instantiated with a concrete mapping (*MDD*

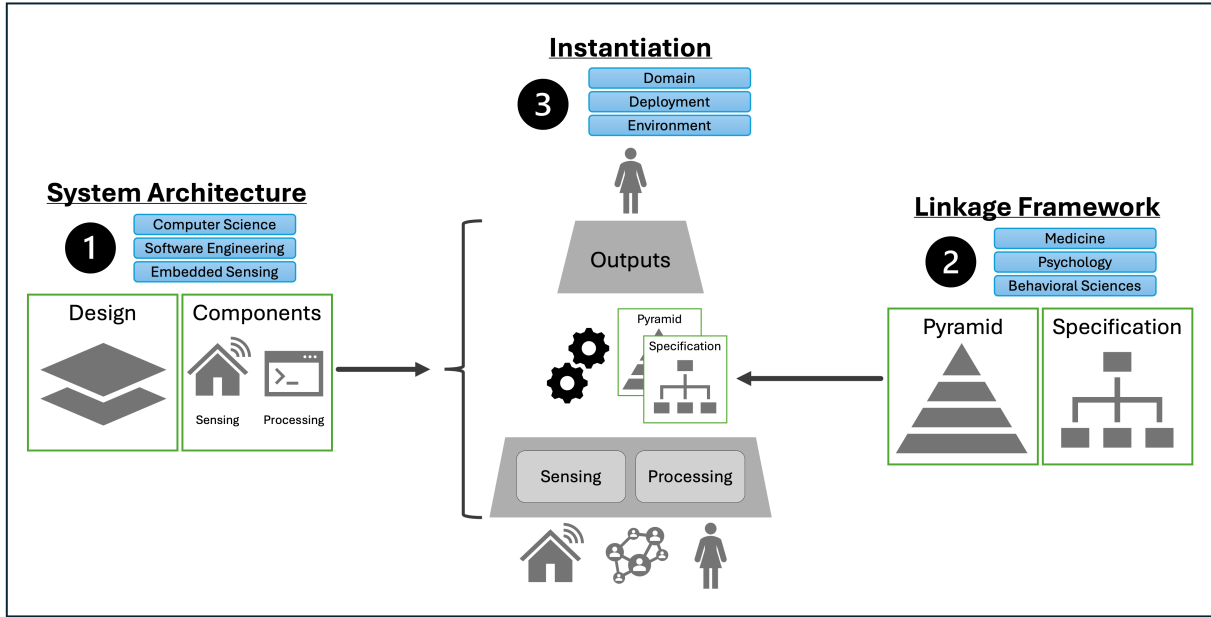


Figure 1.2: Conceptual project idea.

Mapping) for MDD and its corresponding acoustic-based voice features, examining both the degree to which these features reflect depressive symptoms in an attempt to find the most promising candidates for automated acoustic-based speech depression detection.

The second objective of this work is to design a flexible and modular *System Architecture* to support explainable and multimodal automated health assessments (*Obj-2*). This architecture should enable passive, multimodal data collection in IoT-enabled household environments by defining and addressing functional and non-functional architectural requirements. In addition, the system must support exploratory research by allowing easy integration of new modalities, features, and algorithms, and investigate engineering approaches for individual user variability by adapting to diverse user characteristics.

The third objective is to instantiate the proposed *System Architecture* together with the concrete *MDD Mapping* (*Obj-3*). This step presents a software implementation of the proposed *System Architecture* and demonstrates how an instantiation of the *Linkage Framework* can be coherently combined into a functional application.

The last objective is to evaluate both the practical performance of the instantiated *System Architecture* and the practical soundness of the *MDD Mapping* (*Obj-4*). The aim is to assess whether the implementation fulfills the defined architectural requirements and validates the medical relevance of the system's outputs produced by the *Linkage Framework* instantiation.

Table 1.1 shows the formulated objectives.

Table 1.1: Overview of research objectives.

ID	Objective
Obj-1	Design a flexible and modular <i>System Architecture</i> to support explainable and multimodal automated health assessments.
Obj-2	Develop a <i>Linkage Framework</i> that semantically maps medical conditions to biomarkers and instantiate the framework with the mapping for MDD and acoustic-based voice features.
Obj-3	Combine <i>Obj-1</i> and <i>Obj-2</i> . Instantiate the proposed <i>System Architecture</i> design together with the concrete mapping for MDD and acoustic-based voice features.
Obj-4	Evaluate both the practical performance of the instantiated <i>System Architecture</i> design (<i>Obj-3</i>) and the practical soundness of the concrete mapping for MDD and acoustic-based voice features (<i>Obj-2</i>).

1.4 Methodology

This master’s thesis investigates how an explainable white-box system for automated health assessments in the context of IoT-enabled private household environments can be designed and implemented to promote scientific innovation and collaboration. In particular, this work investigates how voice features based on acoustics, alone or in combination with other modalities, can be integrated into the system. The goal is to leverage modular explainable systems to bridge the interdisciplinary gaps between computing and mental health research.

The work follows the well-established Design Science Research (DSR) approach [103], as it aims to address the real-world problem of enabling automated mental health assessment through the iterative design, implementation, and evaluation of conceptual and practical software artifacts [103]. Initial work includes a review of the foundational research that is needed in order to make this master’s thesis self-contained, including the DSM-5 specification, depression, acoustic voice features relevant to depression, and a summary of the most important existing research about multimodal depression detection. Aligned with the DSR methodology, the conceptual artifacts, namely the *Linkage Framework* and the proposed *System Architecture*, are purposeful artifacts to address generalized forms of a given problem type [103]. The *Linkage Framework* addresses the generalization of the problem about unifying the approach for automated mental health assessments, promoting scientific innovation and exchange. On the other hand, the *System Architecture* is a generalized solution for addressing common challenges

faced by professionals, software developers and researchers when working on automated mental health assessment systems. These challenges involve topics around sensing devices, streaming pipelines, distributed architectures, reusable components, multimodal data integration, and more. To address them, the *System Architecture* provides reusable structures and methods that support the efficient instantiation of such systems, enabling knowledge transfer across use cases.

Following the DSR methodology, these artifacts are not only conceptually designed but also instantiated as concrete use case prototypes, *i.e.*, the concrete acoustic-based *MDD Mapping* and a corresponding Python implementation of the proposed *System Architecture* [103].

Aligning with the evaluation phase of the DSR methodology, the evaluation of the two contributed artifacts includes a technical and a conceptual evaluation. The *System Architecture* and its Python implementation are evaluated, where a binary verification of the architecture's ability to meet predefined functional requirements and a qualitative analysis of how well it satisfies non-functional requirements are performed. To evaluate the *Linkage Framework*, its ability to generalize and structure knowledge about mental health conditions is assessed. Its concrete instantiation for the *MDD Mapping* is evaluated based on a set of hypotheses that is rooted within the scientific findings of the literature review, and it is checked whether the hypotheses are valid or not [225], evaluating its practical soundness in efficiently and effectively detecting depression using acoustic voice features.

Chapter 2

Fundamentals

2.1 Background

2.1.1 Emotions

Human emotions are naturally complex. They evolve over time, and each person can express their emotions in a different way. Over the past decades, research activities on emotions and the development of emotion classification models have increased substantially [250].

Shaver *et al.* performed a study where emotions were grouped into prototypes. First, they formed 135 emotions and iteratively minimized them by merging based on their similarity [217]. The end results were the basic (primary) emotions: love, joy, surprise, anger, sadness, and fear [217]. A tree-like structure is used to further classify secondary and tertiary emotions [217]. The study implies that emotions are mixtures of these six basic emotions and that they are mostly indistinct [250]. However, other models exist, such as the tri-dimensional emotion model by Latinjak [129]. This model dynamically categorizes 20 emotion concepts within a cube along the three dimensions of arousal, valence, and time perspective, addressing the issue of overlapping concepts [250], see Figure 2.1. The notion of indistinct emotions underpins the complexity that comes with any attempt to categorize human emotions. Furthermore, the lack of consensus among various emotion models reflects the challenge of capturing the full spectrum of emotional experience within a single framework.

Although this master's thesis focuses on depression in general, which intuitively suggests the association with the basic emotion *sadness*, the challenge of this work is to navigate the multidimensional nature of emotions and to capture these complexities in a

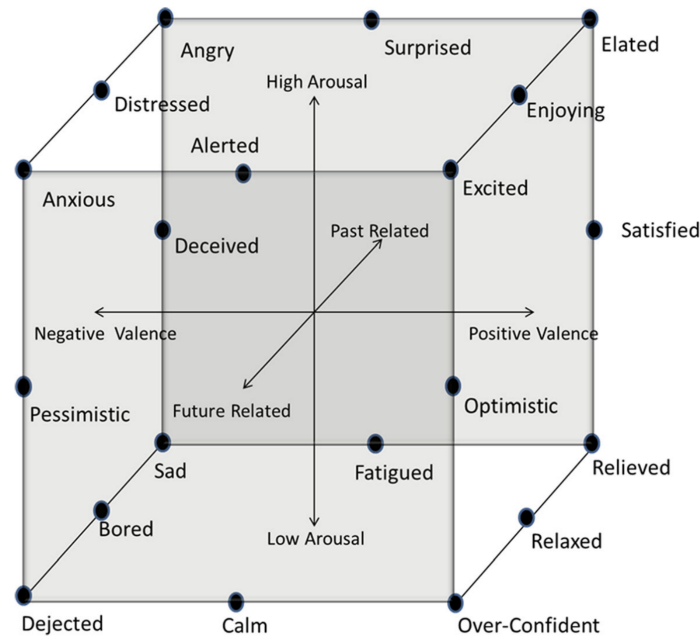


Figure 2.1: Tri-dimensional emotion model by Latinjak [250].

conceptual framework such that it becomes interpretable – much like the field of emotion classification.

2.1.2 Classification Systems

Building on the previous discussion of emotion classification, the focus now turns to the classification of depression. A clinical classification framework is a tool for mental health professionals to identify, describe, and categorize mental illnesses on the basis of distinguishable sets of specific symptoms [135]. Today, the two major classification systems for mental disorders are the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) of the American Psychiatric Association (APA) [18] and the International Classification of Diseases, 11th Revision (ICD-11) by the World Health Organization (WHO) [136]. They are also the most commonly used frameworks [79]. While these frameworks are designed to assist clinicians in conducting clinical assessments, case formulation, and treatment planning, they are also useful for courts and attorneys to assess the legal consequences of mental disorders [18]. Keep in mind, however, that they do not provide treatment guidelines for any given disorder [135]. DSM-5 explicitly states that DSM-5 should not be applied mechanically by individuals without proper clinical training or working experience in the field of psychological diagnostics [18]. Moreover, the case formulation for any given individual should involve

a careful clinical evaluation including social, psychological, and biological factors that may be involved with a given mental disorder [18].

2.1.3 Depression

According to DSM-5, depressive disorders are classified into five main categories: disruptive mood dysregulation disorder, MDD, persistent depressive disorder (dysthymia), premenstrual dysphoric disorder, and depressive disorder due to another medical condition [18].

During the course of this master’s thesis, the later introduced conceptual *Linkage Framework* is instantiated with a concrete mapping for MDD and tone-based voice biomarkers, the *MDD Mapping*. Therefore, further work focuses solely on MDD. The terms *depression* and *MDD* will be used interchangeably in the remainder of this work.

Although each category of depressive disorder is described with its most discriminative clinical indicators, MDD represents the classic condition [18]. MDD is the most reported category of depressive disorders, and more importantly, it is one of the leading causes of disability worldwide and is predicted to be in the first place by 2030 [135]. The most prevalent symptom is anhedonia (the inability to feel pleasure). Common visible physical signs include fatigue, unintended weight loss, and reduced appetite [18, 122]. MDD is often accompanied by low motivation, sleep deprivation, cognitive problems, and feelings of guilt. In other words, MDD affects the individual’s ability to function in their daily life [18, 46]. In addition, for the complex set of symptoms, duration, timing, and etiology may vary between patients [18].

Although great progress has been made in recent years in addressing patients suffering from MDD, including diagnostic methods, medication, and therapeutic methodologies, MDD remains complex and unclear in its nature [122]. Despite its undisputed prevalence, patients deny having depressive feelings in almost half of the cases [160]. Due to false individual perception or stigmatization, nearly 60% of people with depression do not seek medical help [38]. Patients are often brought in for treatment by family members or even sent by the employer to be evaluated. Therefore, due to this widespread denial of their depressive episode, it is crucial to evaluate a patient for suicidality or homicidality at each visit [160]. These circumstances, the alarmingly increasing numbers of incidents, and a dominant situation where, due to false perception or lack of self-reflection, individuals hesitate to seek help, make MDD the ideal candidate for automated depression detection.

The DSM-5 diagnostic criteria for MDD are defined in Table 2.1 [18], and will find application in the next chapters about the design (Chapter 3) and the implementation (Chapter 4) of this work. MDD itself comes in different variations. Those different types of MDD depend on whether this is a single or recurrent episode, the current severity, the presence of psychotic characteristics, and the state of remission [18]. The codes that are included in the MDD codebook range from F32.0 to F32.9 for single episodes and F33.0 to F33.41 for recurrent episodes, see Table 2.2 [18]. While DSM-5 is a widely recognized tool among professionals, it has also faced criticism from some experts. It fails to identify the proper from the correct disorder diagnosis in individual cases. This is due to the fact that patients with the same disorder could suffer from different sets of symptoms [114]. Furthermore, comorbidities or co-occurring conditions pose new challenges to the diagnosis of mental disorders, for which DSM-5 does not account [238]. The taxonomies in DSM-5 are mostly based on various statistical methodologies [18]. Therefore, one of the main shortcomings of DSM-5 is that it lacks culture-specific meaning, *i.e.*, has a clear western world bias while not including the cultural context of certain nations [79]. Lastly, a straightforward diagnosis with DSM-5 could lead to an increase in the number of mentally ill people due to the broadening and generalization mechanism of such a classification framework and therefore do more damage than good through pathological overdiagnosing [29].

Table 2.2: DSM-5 coding for MDD [18].

Severity/course specifier	Single episode	Recurrent episode
Mild	F32.0	F33.0
Moderate	F32.1	F33.1
Severe	F32.2	F33.2
With psychotic features	F32.3	F33.3
In partial remission	F32.4	F33.41

2.1.4 Medical Assessment

In practice, it is difficult to reliably diagnose depression given that symptoms can mimic other disorders or coexisting conditions [111]. A study by Regier *et al.*, called *The DSM-5 Field Trials*, estimated the degree to which clinicians could independently agree on the presence or absence of depression by applying the diagnostic criteria of MDD during

Table 2.1: DSM-5 diagnostic criteria for MDD [18].

Symbol	Description
A	Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure. Note: Do not include symptoms that are clearly attributable to another medical condition.
A1	Depressed mood most of the day, nearly every day, as indicated by either subjective report (<i>e.g.</i> , feels sad, empty, hopeless) or observation made by others (<i>e.g.</i> , appears tearful). (<i>Note:</i> In children and adolescents, can be irritable mood.)
A2	Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation).
A3	Significant weight loss when not dieting or weight gain (<i>e.g.</i> , a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day. (Note: In children, consider failure to make expected weight gain.)
A4	Insomnia or hypersomnia nearly every day.
A5	Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down).
A6	Fatigue or loss of energy nearly every day.
A7	Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).
A8	Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).
A9	Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.
B	The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.
C	The episode is not attributable to the physiological effects of a substance or another medical condition. Note: Criteria A–C represent a major depressive episode. Note: Responses to a significant loss (<i>e.g.</i> , bereavement, financial ruin, losses from a natural disaster, a serious medical illness or disability) may include the feelings of intense sadness, rumination about the loss, insomnia, poor appetite, and weight loss noted in Criterion A, which may resemble a depressive episode. Although such symptoms may be understandable or considered appropriate to the loss, the presence of a major depressive episode in addition to the normal response to a significant loss should also be carefully considered. This decision inevitably requires the exercise of clinical judgment based on the individual's history and the cultural norms for the expression of distress in the context of loss.
D	At least one major depressive episode is not better explained by schizoaffective disorder and is not superimposed on schizophrenia, schizophreniform disorder, delusional disorder, or other specified and unspecified schizophrenia spectrum and other psychotic disorders.
E	There has never been a manic episode or a hypomanic episode. Note: This exclusion does not apply if all of the manic-like or hypomanic-like episodes are substance-induced or are attributable to the physiological effects of another medical condition.

clinical interviews [42]. The results were rather shocking: The inter-rater reliability for MDD was "questionable" (grouping between classes; $\kappa = 0.28$), highlighting the difficulty in finding agreement in how the diagnosis is applied in practice [42]. Kvig *et al.* underlines the difficulty in agreeing on depression diagnosis, reporting also poor consensus results ($\kappa = 0.21$) [54].

Currently, there is no standardized laboratory-based test that is widely used to assess depression; however, many individuals are diagnosed based on how they respond to antidepressant medications [111]. Both traditional clinical interview-based assessment methods, as well as novel non-laboratory and laboratory approaches, exist and have been found useful for depression screening, each demonstrating varying degrees of clinical development and robustness [111].

There are multiple different interview-based methods. For instance, the classic clinical evaluation, which involves a combination of the medical history, mental status examination, and a psychiatric evaluation using diagnostic criteria from standards like DSM-5 [133]. There are standardized diagnostic interviews (SDI) that include structured and semi-structured interview types, such as the most widely used Structured Clinical Interview for DSM-5 Disorder (SCID-5) [133]. The goal of SDIs is an attempt to reduce variability [54]. More than 280 different questionnaires have been developed to assess symptoms of depression in the last century [133]. Clinicians agree that combinations of different methods and tests could be the most effective way to diagnose depression, underlining the importance of a multifaceted (and multimodal) approach to the diagnosis of depression [133]. In practice, clinical proficiency is necessary. In the clinical evaluation, categories such as appearance, behavior, motor activity, speech, mood, affect, thought process, thought content, perceptual disturbances, cognition, insight, and judgment need to be taken into consideration while making a diagnosis [194]. These factors lead to inconsistent results, as an article in the *Psychiatric Times* has shown: Analyzing more than 38 studies and almost 16,000 patients, poor agreement between clinical evaluations and SDI was shown [199]. In about half of the cases, a positive diagnosis of a depression disorder often fails to match SDI findings [20].

In this master's thesis, the quantification of depressive indicators is a central concept, as the severity of depression plays a crucial role in the direction of treatment decisions and subsequent actions. Therefore, measurement of depression severity is a key focus of this investigation.

The PHQ-9 questionnaire, see Appendix A, is used for the feedback mechanism of the proposed *System Architecture* in Chapter 4. It is a useful tool for assessing the severity of depression and can be completed independently by the patient [109]. The nine included questions directly map the DSM-5 indicators on a severity measurement scale from "0 (not at all)" to "3 (nearly every day)" [109]. See Table 2.3 for how the total score of PHQ-9 is interpreted [109]. A study by the creators of the PHQ-9 questionnaire has shown that PHQ-9 achieves a sensitivity and specificity of 88% for MDD [109].

Table 2.3: PHQ-9 total score interpretation for depression severity [109].

Total score	Depression severity
1–4	Minimal depression.
5–9	Mild depression.
10–14	Moderate depression.
15–19	Moderately severe depression.
20–27	Severe depression.

In recent years, the methodology of novel approaches for detecting and diagnosing depression has expanded significantly through the integration of scientific advances in fields such as biology, physiology, genomics, proteomics, and metabolomics [115]. Current advances focus on finding biomarkers for individual diagnostic profiles [181]. As depression is a multifactorial condition, the multimodal approach to a reliable diagnosis of depression seems straightforward [147]. Therefore, diagnostic profiles include biomarkers of different modalities and origins [181]. Although this master’s thesis focuses on expanding the set of useful tone-based voice biomarkers for depression detection, the review first shifts at other directions of biomarker analysis.

Laboratory methods for assessing depression involve biochemical and physiological tests to identify signs of depression, including hormone examinations such as the dexamethasone suppression test (DST) [182]. DST tests how the body responds to dexamethasone, a cortisol-like hormone, which should lower cortisol levels in healthy individuals, while depressed patients show increased cortisol levels after the test [182]. The test lacks good clinical performance for depression assessments [127], yet is capable of distinguishing severe melancholic depression, mania, or acute psychosis from chronic psychosis or dysthymia [72]. Due to the lack of objective physiological indicators and

specific laboratory tests, a reliable diagnosis of depression by biochemical means remains a challenge [236].

Advances in genomics and recent advances in rapid genome sequencing have enabled new possibilities for gene expression profiling, which can be leveraged for the identification of disease classifiers and risk markers in psychiatric disorders [76]. One study identified a seven-gene signature in blood cells that can differentiate between patients suffering from MDD and healthy individuals. The results were confirmed and validated in a replication cohort where the separation of these two groups was found to be statistically significant [215].

These techniques involve the study of different protein profiles in diseased and non-diseased individuals. The goal is to identify different profiles of metabolites between groups of diseased and non-diseased individuals [121]. These metabolites then can be identified as profiles that then are expected to be able to identify depressed patients [121]. By exposing a patient to antidepressant treatment, alterations in energy and the phospholipid metabolism as well as Ca^{2+} homeostasis are investigated [13, 51, 213]. Various studies show that antidepressants alter neurogenesis-related markers such as the brain-derived neurotrophic factor (BDNF) [146]. While already promising techniques have been found, further exploration is needed to validate existing findings [115].

A recent study by Liu *et al.* highlights the importance of further biomarker research and emphasizes the potential to improve the general understanding of depressive disorders while improving current diagnostic methods [235]. This object is at the core of this master's thesis (*Obj-2*, see Table 1.1). This work aims to expand existing research on depression biomarkers by focusing on tone-based voice biomarkers that contribute to the reliability and precision of a holistic approach to depression detection. It is evident that tone-based voice biomarkers hold great potential for identifying depression, as various other biomarkers associated with depression can influence vocal characteristics [235]. The hypothalamic-pituitary-adrenal (HPA) axis regulates cortisol, the main stress hormone, and affects vocal characteristics such as pitch, intensity, rhythm, and pace [235]. Cytokines and the inflammatory response affect brain regions responsible for motivation, energy, and motor control, which can result in low-energy speech, altered speech rate, and increased pause frequency [235]. Neurotrophins, which play a crucial role in emotion regulation, are also often found at reduced levels in depression patients, leading to diminished pitch variability and reduced emotional expressiveness in speech. Genetic and epigenetic factors have also been shown to contribute to vulnerability to

depression [132]. Individuals with similar genetics may exhibit similar patterns in vocal prosody [235].

2.1.5 Tone-based Approach

Human social behavior has contributed significantly to the understanding of emotions. Research has shown that people from diverse cultures can reliably identify emotions from facial expressions [179]. Similarly, listeners are often able to infer emotional states based on prosodic voice characteristics such as tone, pitch, and rhythm [217].

Speech carries a lot of data. Speech Emotion Recognition (SER) is a discipline that uses these data to automatically classify speech into distinct emotional categories [140]. Speech data consists of both acoustic data (the "how") and linguistic data (the "what") [26], see Figure 2.2. While knowledge derived from linguistic data is largely objective, extracting information from acoustic data inherently involves a degree of subjectivity [26]. This subjectivity can pose challenges, particularly in clinical depression assessments based on interview methods. Therefore, this work attempts to propose a tone-based speech approach for the detection of depression.

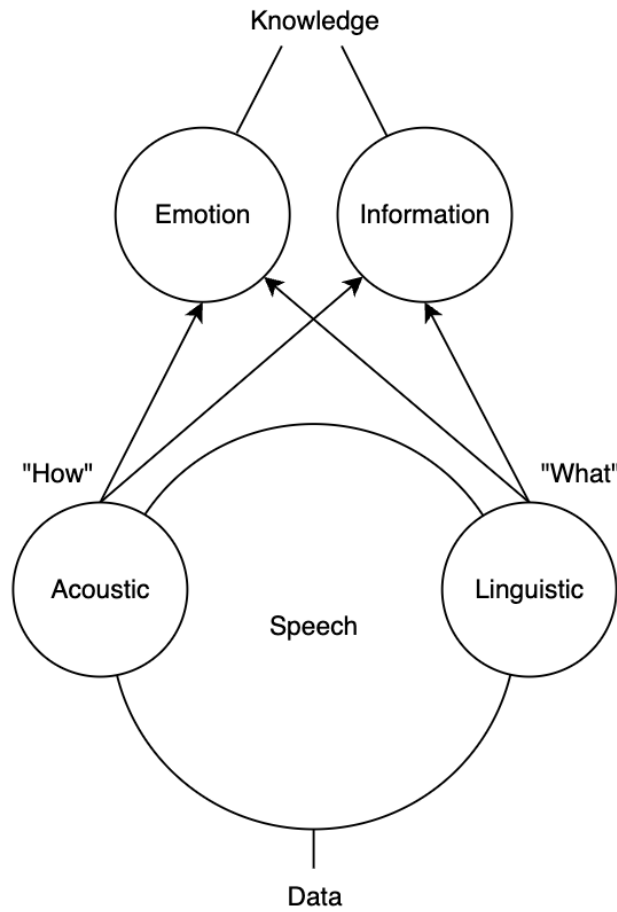


Figure 2.2: Bimodal concept of knowledge in speech data [26].

To address these issues, this master’s thesis employs passive and continuous audio sensing within the private household setting, which allows collection of naturalistic vocal behavior. This approach reduces the problem of the patient altering its voice or following certain self-presentation strategies. Furthermore, this approach may reveal even stronger characteristics in vocal biomarkers than in the case of structured interviews or laboratory studies [70]; also see Section 1.2.

It is important to understand the production of speech. Figure 2.3 illustrates that this process consists of two parts: the cognitive planning and the motor action [162].

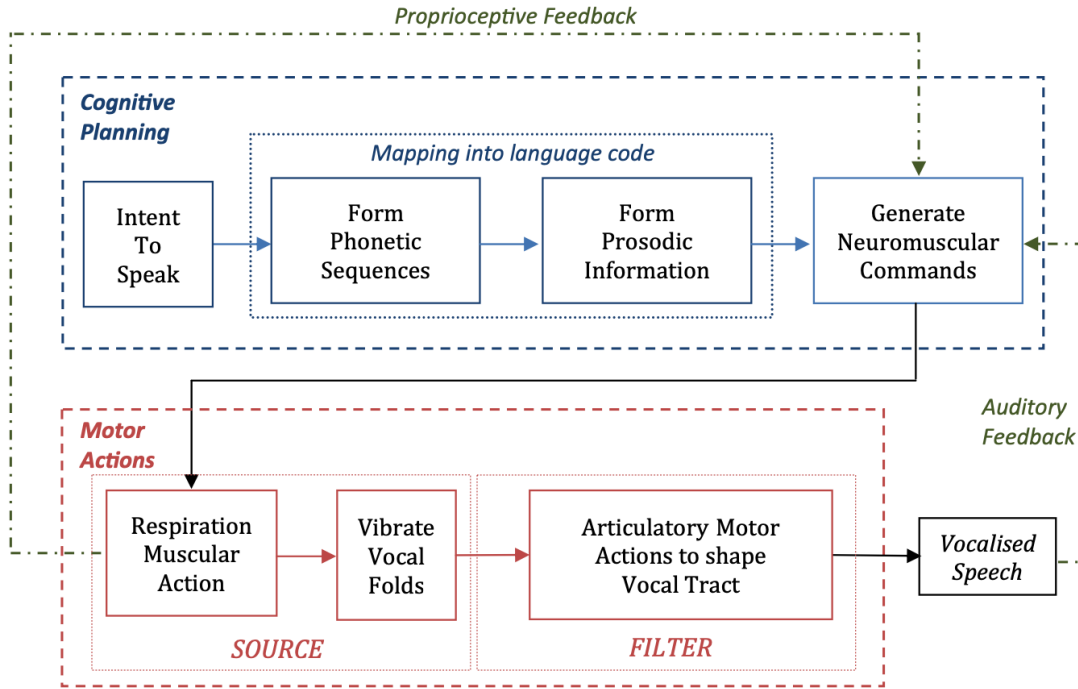


Figure 2.3: Speech production process [162].

For this acoustic-based approach, the motor action is particularly interesting. Voice is produced through a coordinated process involving the lungs, vocal cords, vocal tract, nasal cavity, and lips [162]. The airflow is compressed from the lungs and reaches the vocal cords through the trachea. The airflow generated by the vibration of the vocal cords resonates in the vocal cavity, producing sound waves and thus forming an acoustic signal [162]. Sound waves are a form of waveforms that change continuously over time and are a continuous (analog) signal. Since computers cannot process analog signals directly, the voice signal must be sampled and converted into a digital form for analysis and processing [245].

2.2 Related Work

2.2.1 ML-based Depression Detection

Since the rapid advancement of ML and AI, the use of ML algorithms has penetrated various areas in our world [248]. While these algorithms are widely adopted in health-care and medical domains, their practical and clinical use in the field of psychology remains limited [112]. In certain disciplines of health care, ML approaches already outperform humans. One such example is the Google Inception v3 model [30], which identifies diabetic retinopathy from photographs [66]. Many different ML approaches

and architectures have been successfully used for automated depression detection. Neural networks (NN) are particularly well suited to this task, as they can effectively model complex, multidimensional patterns associated with depressive symptoms [105, 229, 232]. However, ML models are black boxes; they do not provide any explanation or reasoning for why a certain label has been associated with the given input [66]. The model only knows that there exists some kind of correlation between label and input [248]. Furthermore, only a few researchers formulate what interpretability means and how it is implemented in their systems [248]. The inability of us humans to understand how these ML models work poses a great risk and leads to two issues: Research indicates that a lack of trust in ML models — particularly in the context of depression diagnosis — is a significant barrier to their adoption in the medical field [112]. Additionally, the lack of transparency in these black-box models limits our understanding of how they arrive at their predictions, offering little guidance for treatment decisions or other critical aspects of medical care [112]. Current research explores how decision-making processes in ML models can be made transparent and more self-explanatory to mitigate the above issues and improve the practitioner’s understanding of automated depression detection and mental health decision-making in general [112].

2.2.2 Modality

Previous studies have primarily focused on text, audio, and video modalities for automated depression detection [112]. Each of these high-level modalities can be further broken down into specific data types originating from various sources, focusing on individual aspects, such as social media, messaging services, speech, vocal characteristics, body posture, facial expression, etc. [226]. As for this master’s thesis, the setting foresees an IoT-enabled private household. This promotes the utilization of various different sensors and their unique modality, which potentially have not yet been investigated in research. The possibilities are limitless, ranging from walking patterns within the home environment, sleeping patterns, and water- and food consumption to physiological signals from smart devices such as weight fluctuations or heart rate variability.

2.2.2.1 Text

Extensive research has been dedicated towards the analysis of textual content from social media networks (SMN), see Table 2.4 [60, 71, 176, 189]. As one of the largest sources of unstructured data, it is a promising source for gaining insights into the general public opinion, social behavior, as well as mental health status. The study by Shen

et al. marks a milestone in automated depression detection. They created a well-labeled Twitter dataset and were able to detect depression with high accuracy [71]. Their model identified additional behavior meta-information types that were discriminative for depression / non-depression, such as posting time, emotion catharsis live sharing, etc. [71]. Chiong *et al.* built on this by improving the generalization of depression detection, which is a key limitation of the work by Shen *et al.* [71]. By eliminating explicit words that are associated with depression, such as "depression" or "diagnosis", they were able to detect more subtle depression / non-depression patterns hidden in the texts [189]. Furthermore, they tested their model using text from nonsocial networks, thus validating the generalization [189]. Another aspect of automated depression detection is the system's time awareness. Cacheda *et al.* developed a time-aware system that rewards early detections and penalizes late detections, which has the potential of improving detection methods by more than 10% [60].

Having this fragmented work of psychological disorder detection (PDD), Nalinde *et al.* proposed a framework that unifies the PDD process in the three main components: collecting data, feature extraction, and classification [176]. Furthermore, during the collection of data and the feature extraction steps, it is important to highlight the importance of appropriate preprocessing and extraction techniques. For instance, Chiong *et al.* used components from the Natural Language Toolkit [203] for punctuation and number removal, tokenization, stopword removal, and lemmatization, as well as Peter Norvig's code for spelling correction, elongation word correction, and negative word processing [172, 189].

In the setting of depression detection using social media data, Mao *et al.* collected various markers (including metadata) that are positive (use of swear words, time spent on social media), neutral (word count, word embeddings, bag of sub-emotions, BERT embedding) or negative (avg. number of words, number of mentions, update frequency, use of first-person pronouns) correlated with depression [112]. Furthermore, their work presents a wide range of text-based depression detection studies [112].

Table 2.4: Text-based depression detection studies.

Work	Dataset	Methods	Result	Note
Shen <i>et al.</i> , 2017 [71] – Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution	Three individual Twitter datasets: depressed, non-depressed, unlabeled	MDL	85% F1-Measure	Solid baseline approach with clear indicators in meta-data of Tweets for depressive behavior.
Chiong <i>et al.</i> , 2021 [189] – A textual-based featuring approach for depression detection using machine learning classifiers and social media texts	SMN datasets by Shen <i>et al.</i> , 2017 [71] and Eye, 2020 [59]; additional non-SMN datasets from Kaggle (Victoria’s Diary, Reddit, Facebook)	Linear Regression, LSVM, Multilayer Perceptron, Decision Tree	LSVM (highest): 99.81% F1-Measure	Framework for preprocessing techniques; generalization of depression detection models.
Cacheda <i>et al.</i> , 2019 [60] – Early Detection of Depression: Social Network Analysis and Random Forest Techniques	eRisk challenge dataset, 2017	Two Random Forest classifiers	>10% ERDE50 performance	Applied temporal-aware evaluation.
Nalinde <i>et al.</i> , 2019 [176] – Machine Learning Framework for Detection of Psychological Disorders at OSN	Custom dataset consisting of Facebook posts	Naive Bayes, SVM	79.9% accuracy	Proposed ML framework for depression detection.

2.2.2.2 Audio

Over the past decades, depression detection using audio data has been studied intensively and has emerged as a promising approach, see Table 2.5. The paper of Cummins *et al.* was the first review into the analysis of speech for the detection of depression and suicidality [162]. One important contribution of their paper is the review on depressed and suicidal speech databases that have had results published on them in the last ten years [162]. Another contribution of that paper is the summary of benchmark value distributions for a wide set of acoustic voice features [162]. Furthermore, they identified six important factors that need to be considered when choosing the ideal set of acoustic voice metrics [162]. As this master’s thesis revolves around the audio-based approach, not only similar studies but also concrete voice features are discussed in the following paragraphs. Most of the existing research has focused on speech analysis approaches, aiming to identify vocal biomarkers associated with depressive behavior. Already in 1921, Kraepelin described depressive voices as those of patients speaking in a low and slow voice. He notes that

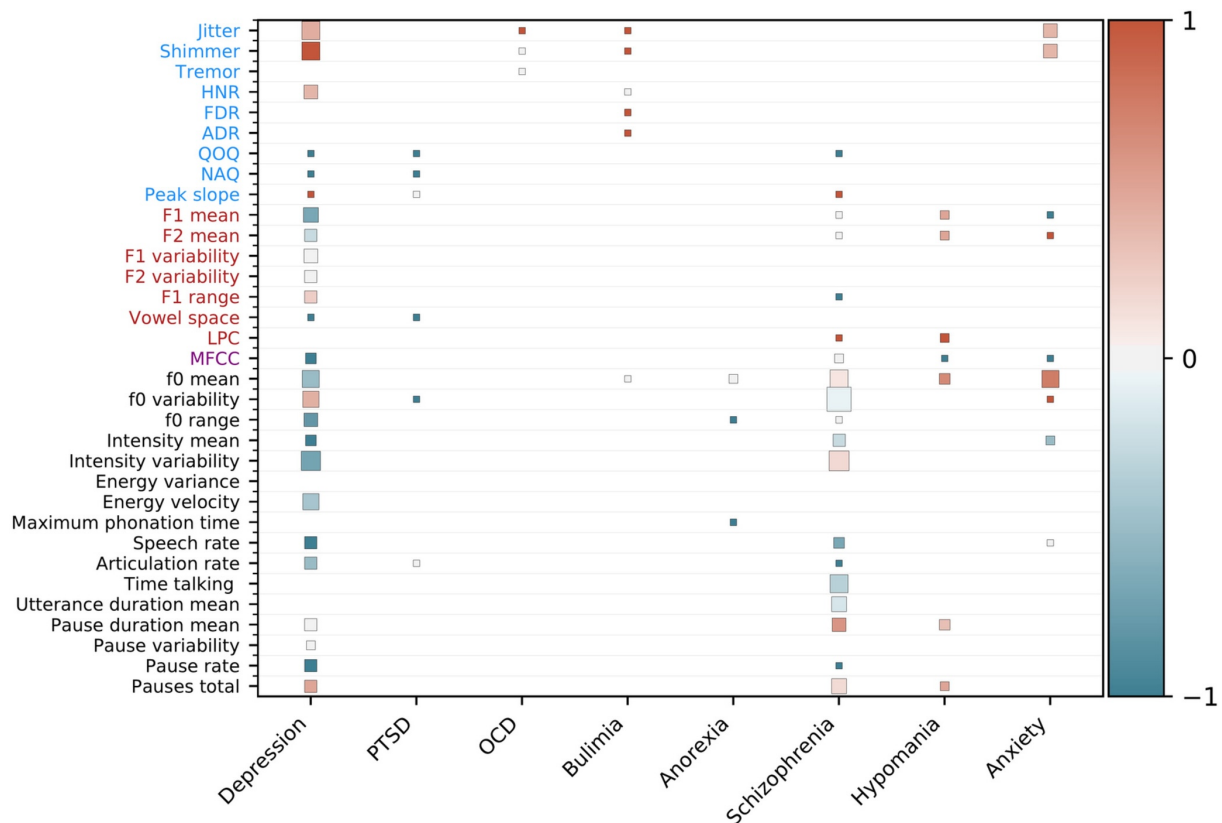


Figure 2.4: Voice feature effect sizes according to Low *et al.* [50].

such speech is often characterized by hesitation, monotony, whispering, and occasional stuttering. Furthermore, patients may abruptly stop mid-sentence [120].

An early work by Cohn *et al.* compared depression with affective computing and found that particular vocal biomarkers, such as voice prosody, can offer valuable insights into the individual’s mental health state. They leveraged pitch extraction methods to measure vocal expression and logistic regression classifiers, resulting in an overall 79% classification accuracy [93].

Voice features can be divided roughly into the four groups: source features, spectral features, prosodic features, and formant features [50]. Figure 2.4 summarizes findings from null-hypothesis testing studies across psychiatric disorders, highlighting which acoustic features differ significantly between clinical and control groups. Source features (blue), spectral features (purple), prosodic features (black), and formant features (red) with effects scored from -1 (negative correlation) to +1 (positive correlation) [50].

A similar study by Di *et al.* investigated a set of 20 pitch-related acoustic voice features and strongly associated eleven features with depressive symptoms, such as a slower pitch change and a lower pitch [239]. In addition, the review paper by Mao *et*

al. provides an aggregated overview of the most commonly used voice features for depression detection [112]. 19 articles from January 2019 to February 2022 explored voice biomarkers for depression using ML, 17 using classification and two using regression methods. The average F1-measure for these studies is 76% [178]. The highest achieved F1-measure was achieved by Zhao *et al.* with 99% [244], while it is important to note that some papers relied on pre-existing models or datasets, which may have benefited from prior insights, potentially influencing the F1-measure scores [178].

While tone-based voice analysis seems to be the most promising audio-based approach, there exists sparse literature that investigates biomarkers beyond vocal features. For instance, Wu *et al.* proposes volume and the zero-crossing rate (ZCR), the number of times the audio passes through zero in each sound frame, as important features in depression detection [245]. They argue that individuals with depression are more likely to produce speech patterns where noise and breath sounds exhibit a higher ZCR than typical speech [245]. Adding to this complexity and many different sub-disciplines of a tone-based depression detection approach, it has been shown that the speech productions are either externally or unconsciously affected by many factors. For example, a change in the vocal cords has been reported after a COVID-19 infection and therefore changes in vocal biomarkers [64].

The work by Huang *et al.* represents another great milestone. They used *wav2vec 2.0* as a voice-based pre-training model for the feature extraction [232]. The speech embeddings then were used in a classification model and achieved on the popular DAIC-WOZ dataset [41] an outstanding classification accuracy of 96.49% [232].

Table 2.5: Audio-based depression detection studies.

Work	Dataset	Methods	Result	Note
Cummins <i>et al.</i> , 2015 [162] – A review of depression and suicide risk assessment using speech analysis	No dataset	Literature review	Identified six factors for selecting acoustic voice features. Benchmark value distribution for a set of acoustic voice features.	Summary review of speech analysis depression detection studies and speech databases.
Di <i>et al.</i> , 2024 [239] – Unraveling the associations between voice pitch and major depressive disorder: a multisite genetic study	Custom voice feature dataset of a multisite case study of Chinese women with recurrent depression.	Used <i>OpenSMILE</i> for extracting the ComParE2016 feature set. Analyzed correlations and statistical significance (p-testing) of voice features	Eleven statistical significant voice features for depression detection, such as slower pitch changes and lower pitch (AUC-ROC: 0.90, sensitivity: 0.85, specificity: 0.81).	They only focused on Chinese women and highlighted that the results may not apply for the male population.
Low <i>et al.</i> , 2020 [50] – Automated assessment of psychiatric disorders using speech: A systematic review	No dataset	Literature review, following the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) guidelines.	Curated an online database with study results and synthesized acoustic voice feature sets for different DSM-5 disorders.	Summarized studies of the last 10 years of research, where 63% of studies used ML approaches and 37% of studies performed null hypothesis testing methods.
Wu <i>et al.</i> , 2021 [245] – Emotion Detection by Speech and Voice Tone	Custom word corpus for Mandarin language.	Implementation work focusing on speech content and the tone change analysis of the voice.	Developed a mobile application that is able to track and detect negative emotions through the voice.	Strong focus on the technical implementation and development of the application.
Huang <i>et al.</i> , 2024 [232] – Depression recognition using voice-based pre-training model	DAIC-WOZ dataset	<i>Wav2vec 2.0</i> voice-based pre-training model as feature extractor and a small fine-tuning classification network.	Accuracy: 96.49% with RMSE of 0.1875 on the test split.	High-level classification of depression and non-depression. Strong in MDD detection with high generalization abilities.

2.2.2.3 Video

In this category, the video modality includes only studies that solely investigate visual data, excluding auditory information, see Table 2.6, which would constitute a multimodal approach. Depression detection based on facial expression is a sub-discipline within the well-established field of Facial Emotion Recognition (FER). For example, Tufail *et al.* employed a convolutional neural network (CNN) to classify facial images into seven basic emotion categories [77]. They raised the important question of whether such a

lab-based classification approach, limited to predefined emotional classes, could translate effectively to real-world depression detection scenarios [77]. Kong *et al.* partly addressed this question by specifically performing depression detection of portraits. They used different types of CNN architectures and achieved the highest F1-measure of 98.27% with a fully convolutional CNN (FCN) [234]. Other physical cues expressed by the body can also indicate depression and be captured via a camera. Liu *et al.* combined facial expression and body posture features extracted from video data to detect depression, achieving a maximum F1-measure of 90.1% using a decision-level fusion model [241].

Also in the setting of depression detection using facial expression, Mao *et al.* again summarizes a collection of beneficial markers that are neutral (local binary pattern) or negatively (right nose speed, points at the corners of the mouth, distances across left and right eye brows, pupil movement, AU15, AU14, AU12) correlated with depression [112]. Again, they include supplementary results from related work on video-based depression detection [112].

Table 2.6: Video-based depression detection studies.

Work	Dataset	Methods	Result	Note
Liu <i>et al.</i> , 2025 [241] – Multimodal depression recognition and analysis: Facial expression and body posture changes via emotional stimuli	Proprietary experiment video data: patient group (73), control group (73)	ResNet50, decision-level fusion model	90.4% accuracy; 90.1% F1-measure	Combination of facial expression and body posture features.
Kong <i>et al.</i> , 2022 [234] – Automatic Identification of Depression Using Facial Images with Deep Convolutional Neural Network	Facial photos dataset by Department of Neurology, Affiliated Hospital of Shandong University of Traditional Chinese Medicine (Oct 2021–May 2022)	FCN, Vgg11, Vgg19, ResNet50, Inception-V3	FCN (highest): 98.27% F1-measure	Solid baseline approach for facial depression detection using CNNs.

2.2.3 Multimodality

For a core contribution of this master’s thesis, this work proposes a *Linkage Framework* and a *System Architecture* that foresee the integration of multiple data streams of varying modalities. Most existing approaches rely predominantly on unimodal data streams such as text, audio, images, or video [192]. In recent years, multimodal approaches have gained significant attention in affective computing. Since no single ML algorithm consistently outperforms others in all tasks, it becomes crucial to select the most suitable model for

each specific problem. Multimodal methods address this challenge by taking advantage of the complementary strengths of different data modalities, which often leads to more robust and accurate performance [49]. Recent studies demonstrate the possible benefits of a multimodal approach, particularly in this setting of mental health assessment.

Min *et al.* have shown in their work that the combination of audio and imagery data can substantially improve the accuracy of depression detection. Each input data stream mode contributes in its own unique way to the final assessment. For instance, they have found that intensity of voice (audio) and smile frequency (imagery) are significant predictors and produce good results in depression detection [113]. Another study by Sadeghi *et al.* pursued the development of a ML model following a multimodal approach with text-based features using a large language model (LLM), facial features using a bidirectional LSTM network, as well as a combination of both models. They underscored the potential of the multimodal analysis approach by showing the improvements in terms of robustness and effectiveness [151].

The study of Atmaja *et al.* investigated a combined approach of the acoustic and linguistic features for automated SER. By this fusion of both streams of information, they were able to improve the understanding, classification, and performance of SER systems [26].

Video, as a combination of auditory and visual modalities, forms a particularly powerful and dominant approach. Lin *et al.* designed a system called *SenseMood* that demonstrates how depression can be detected using a deep visual-textual multimodal learning approach. They used a CNN-based classifier together with BERT to extract features and combined them using a NN for the analysis [33]. The work by Sharma *et al.* went one step further and quantified the detected emotion of users by assigning scores. Although they still rely solely on textual data from Twitter, they suggested integrating additional data like geospatial information, pictures, sounds, articles, multimedia, etc. for a multimodal approach [12].

From these insights, it becomes evident that many steps of the automated depression detection pipeline are similar or even identical across studies. While they often rely on the same pre-processing and feature extraction techniques, they differ in the employed analysis method or the specific downstream tasks. Furthermore, researchers explore many different approaches to using the different data modalities, highlighting the need for a flexible architecture that facilitates their modular combination and easy data collection.

Chapter 3

Design

This chapter is dedicated to the work on *Obj-2*, see Table 1.1, the design of a *Linkage Framework* that systematically connects medical conditions with relevant biomarkers. To illustrate the framework, it is instantiated with a concrete mapping between the MDD specified by DSM-5 and the respective biomarkers of the acoustic voice.

3.1 Structured Approach

As discussed in Section 2.2, in recent years, the most dominant approach for detecting depressive behaviors has relied on deep learning, ML techniques employing NNs. The models are trained on large datasets of various data types, such as speech, text, or imagery, to reflect patterns that correlate with depressive behaviors. Due to the complexities of the large variations in depression profiles, these ML approaches enable automated depression detection in an efficient and effective way [162]. Among these techniques, CNNs, residual neural networks (RNNs), and NNs are used to tackle the different tasks involved, such as feature extraction, temporal context modeling, or classification [183].

There exists a substantial trade-off between model performance and their explainability: ML models are capable of revealing undiscovered non-linear relationships and latent structures in data that are not necessarily comprehensible by humans [126]. Furthermore, this trade-off is situational depending on the complexity of the data [126]. Although the new potential of training billions of individual parameters is a chance to reveal newly hidden patterns, in the scenario of depression detection, it also introduces the difficulty of reasoning about the clinical analysis. Understanding symptoms, patient history, contextual factors, and the underlying root causes of depression is an essential prerequisite to appropriately prescribe meaningful therapy measures or medications for individual patients [82]. Ultimately, a system that only detects depressive behavior without insight into its reasoning fails to support psychiatric decision-making [82].

These issues underpin the need for an automated depression detection strategy that promotes interpretability. As Fried *et al.* suggests, moving away from total depression scores to understanding symptoms at the symptom level is essential for appropriate treatment [82]. Aligning with their perspective, this master’s thesis provides a design for a structured and interpretable *Linkage Framework* for mapping diagnostic indicators described by DSM-5 with the respective biomarkers to gain insight into the given DSM-5 disorder and to guide appropriate treatment.

3.2 Linkage Framework

3.2.1 Conceptual Idea

The core idea behind the proposed *Linkage Framework* is to address the challenging and complex field of scientific knowledge associated with mental health disorders. At the center of this framework lies the structured relationship between high-level mental health disorders as defined in DSM-5 and the corresponding low-level biomarker data. DSM-5 shows a consistent structure across all the defined disorders, where each disorder is characterized and specified through a defined set of clinical indicators [18]. This opens up the possibility for the abstraction of common patterns. In this way, an automated mental health diagnosis framework is proposed that promotes interpretability, uniformity, and interchangeability.

The interpretability of the framework is enabled by the transparent mapping from empirical, low-level biomarker data to the corresponding DSM-5 clinical indicators. Therefore, professionals can gain a deeper understanding of individual disorder indicators, which could help in the overall evaluation of the disorder and ultimately also the appropriate treatment methods. Because different DSM-5 disorders sometimes share identical clinical indicators and symptoms [18], established mappings for specific indicators can be reused. Through this structured approach, future research on the exploration and investigation of biomarkers associated with DSM-5 disorders can be carried out uniformly and streamlined by facilitating knowledge transfer between different disorder mappings.

In this design, the *Linkage Framework* is instantiated with the high-level DSM-5 disorder MDD, using scientifically validated acoustic voice features as primary low-level biomarkers. Importantly, the framework is agnostic to the modality of biomarkers, which also promotes a holistic and multimodal approach to automated mental health

diagnosis. This conceptual idea of the framework serves as a logical bridge between the disciplines of clinical psychology and computer science. It demonstrates the potential of interdisciplinary integration by allowing computational systems to reason over clinical knowledge while remaining grounded in empirical data.

3.2.2 Mapping Pyramid

The conceptual mapping between high-level DSM-5 disorders and low-level biomarker data is displayed as a multistage network. Decomposing the disorder from top to bottom into increasingly fine-grained classes facilitates knowledge transfer on each layer. This master thesis proposes for the conceptual mapping the model of a pyramid, see Figure 3.1. The foundation of the pyramid consists of the raw data, where the roof of the pyramid represents the final diagnosis of a DSM-5 disorder as the aggregated result of the mapping pyramid.

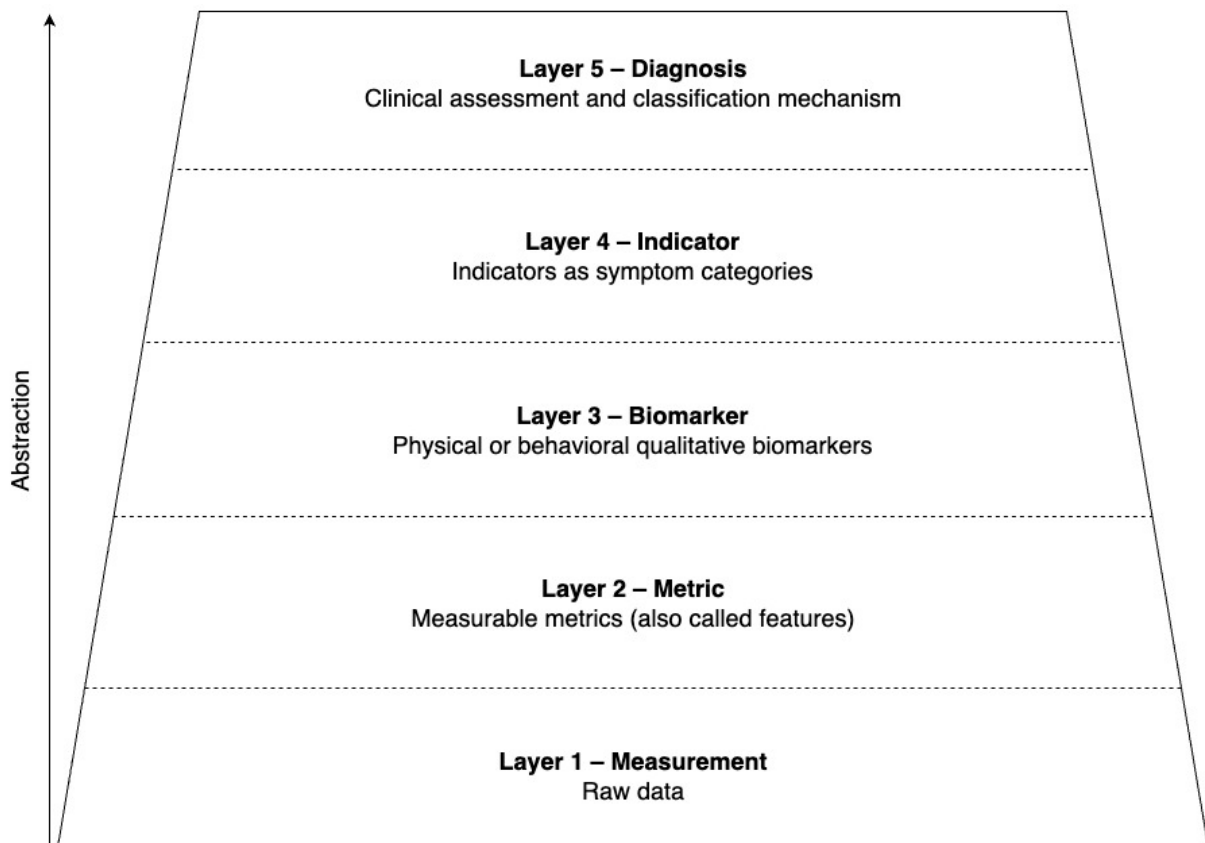


Figure 3.1: *Linkage Framework, conceptual Mapping Pyramid.*

Layer 5 – Diagnosis: At the top of the pyramid is the high-level diagnosis of mental health disorders. In the context of this conceptual mapping, this refers to any mental

health disorder as defined in DSM-5. This layer represents the clinical outcome based on the findings of lower-layer observations and is the final stage of the inference chain. The DSM-5 specification defines clinical indicators and additional mechanisms on how to interpret indicators. These mechanisms are not reflected in the *Linkage Framework*; rather, the clinical professional is in charge of interpreting the indicators, see DSM-5 [18].

Layer 4 – Indicator: This layer is related to the specific clinical indicators of DSM-5 [135]. These indicators are well-defined and uniquely associated with the corresponding DSM-5 disorder of choice. These indicators serve as a guideline for interpreting patient symptoms and behavior [135] and build the logical bridge between symptoms and the DSM-5 disorder. This enables the interpretation of the symptoms in the event of a potential present DSM-5 disorder.

Layer 3 – Biomarker: Still in the realm of clinical psychology, each DSM-5 indicator is linked to a unique set of qualitative biomarkers. They are determined by observing patterns in the individual’s behavior and ideally supported by scientific findings from literature [233]. These biomarkers enable easier detection of the corresponding abstract indicators by building their concrete character profile. Biomarkers can either be identified by human or subjectively described and detected by an automated system [150].

Layer 2 – Feature: The connection between biomarkers and metrics (features) is the glue between the realm of medicine, psychology, and behavioral science and the broader field of computer science, as it forms a transparent mapping from qualitative data to quantitative data [34]. The qualitative biomarkers are evaluated based on objective and measurable features [34]. These features are obtained by aggregating raw data, and the utilized techniques highly depend on the modality at hand.

Layer 1 – Measurement: At the foundation of the *Mapping Pyramid* are the raw data that have been extracted from various sources. In this layer, the preprocessed and error-free data is provided to upper layers to enable straightforward feature computations upon them. Note that the quality of the measurements determines the performance, robustness, and reliability of the overall framework.

3.3 MDD Mapping

To complete *Obj-2*, the proposed *Linkage Framework* is instantiated in the *diagnosis layer* with MDD as defined in the DSM-5 [18].

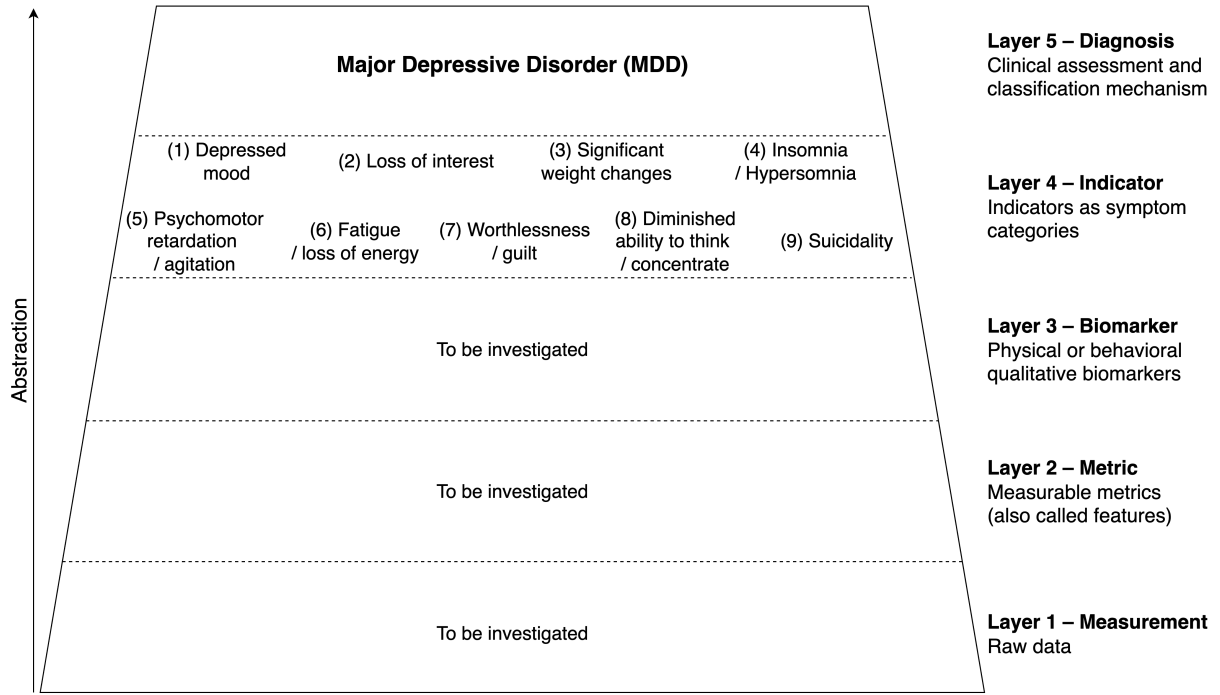


Figure 3.2: Mapping Pyramid instantiated with MDD in the diagnosis layer.

The primary focus of the proposed instantiation lies in finding biomarkers for an audio-based depression detection system. As established in Section 2.1, audio-based depression detection is bimodal in nature. For the search for promising biomarkers, this master’s thesis focuses only on the acoustic stream of data. The concept of the concrete *MDD Mapping* is a promising solution for the research on acoustic-based voice features for depression detection and will help in understanding the connection between voice characteristics and depression.

Figure 3.2 illustrates the top two layers of the instantiated *MDD Mapping*. In the following work, the short form of the clinical indicators is used: (1) *Depressed mood*, (2) *Loss of interest*, (3) *Significant weight changes*, (4) *Insomnia / hypersomnia*, (5) *Psychomotor retardation / agitation*, (6) *Fatigue / loss of energy*, (7) *Worthlessness / guilt*, (8) *Diminished ability to think / concentrate*, (9) *Suicidality* [18]. In simple words, the design approach now consists of finding appropriate acoustic-based voice features for the other layers of this pyramid; see Section 3.4.

3.4 Feature Selection

The following section discusses individual DSM-5 indicators for MDD and how they can be observed through qualitative biomarkers and quantified by acoustic-based voice

features. These acoustic-based voice features may be obtained through algorithmic computation or ML-based techniques [232]. There are substantial overlaps between individual indicators and their respective mappings.

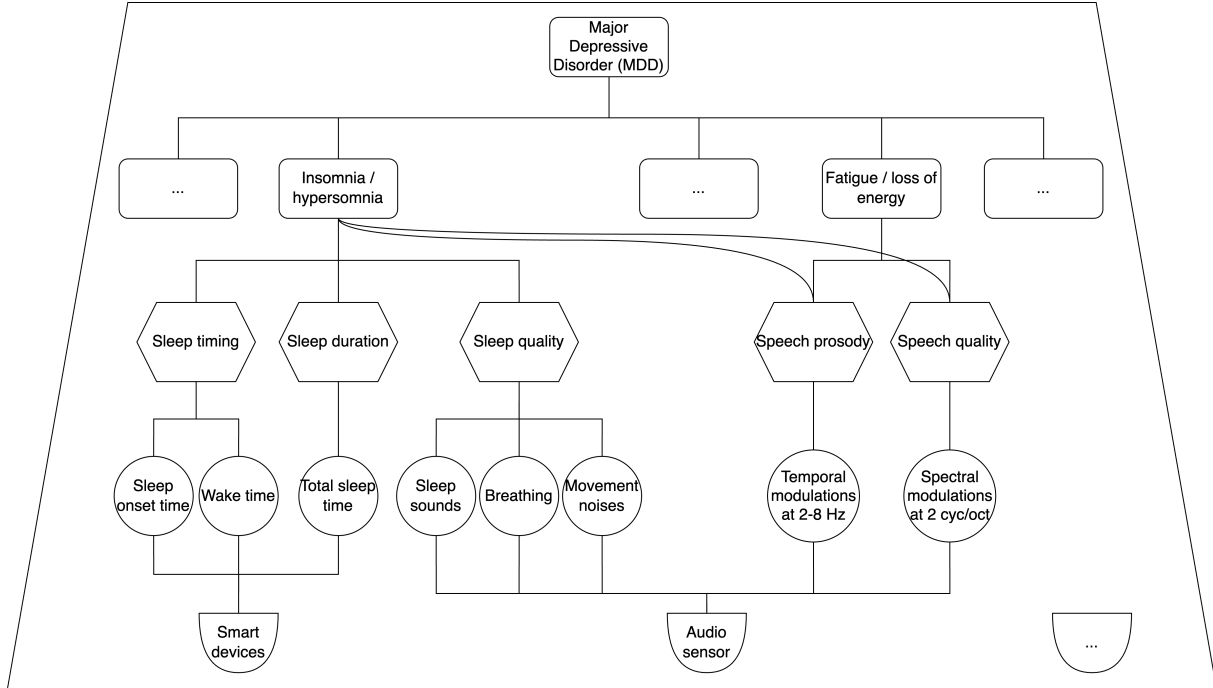


Figure 3.3: Mapping overlap between indicator (4) *Insomnia / hypersomnia* and (6) *Fatigue / loss of energy*.

For example, Figure 3.7 shows an overlap in the biomarkers of the two indicators (4) *Insomnia / hypersomnia* and (6) *Fatigue / loss of energy* mappings. The concrete mapping pyramid will be decomposed in the upcoming sections which is an essential contribution of this master’s thesis and builds the foundation for the practical implementation in Chapter 4.

3.4.1 Depressed Mood

Figure 3.4 illustrates the proposed mapping for the MDD clinical indicator (1) *Depressed mood* [18]. Depressed mood is often audible and is characterized by biomarkers such as whispering, monotonous, slow, or slurred speech, prolonged pauses, or speech quality [104]. While whispering, the vocal cords do not vibrate; therefore, a shift of lower energy formant frequencies, especially in the F1-formant ($F1$) and the F2-formant ($F2$), to higher frequencies and a reduced signal-to-noise ratio (SNR) indicate whispered speech [75, 107, 221]. Monotonous speech is detectable by a reduced fundamental frequency average

($F0_{avg}$), a reduced fundamental frequency standard deviation ($F0_{std}$) [63, 162], and an increase in spectral flatness (SF) [125].

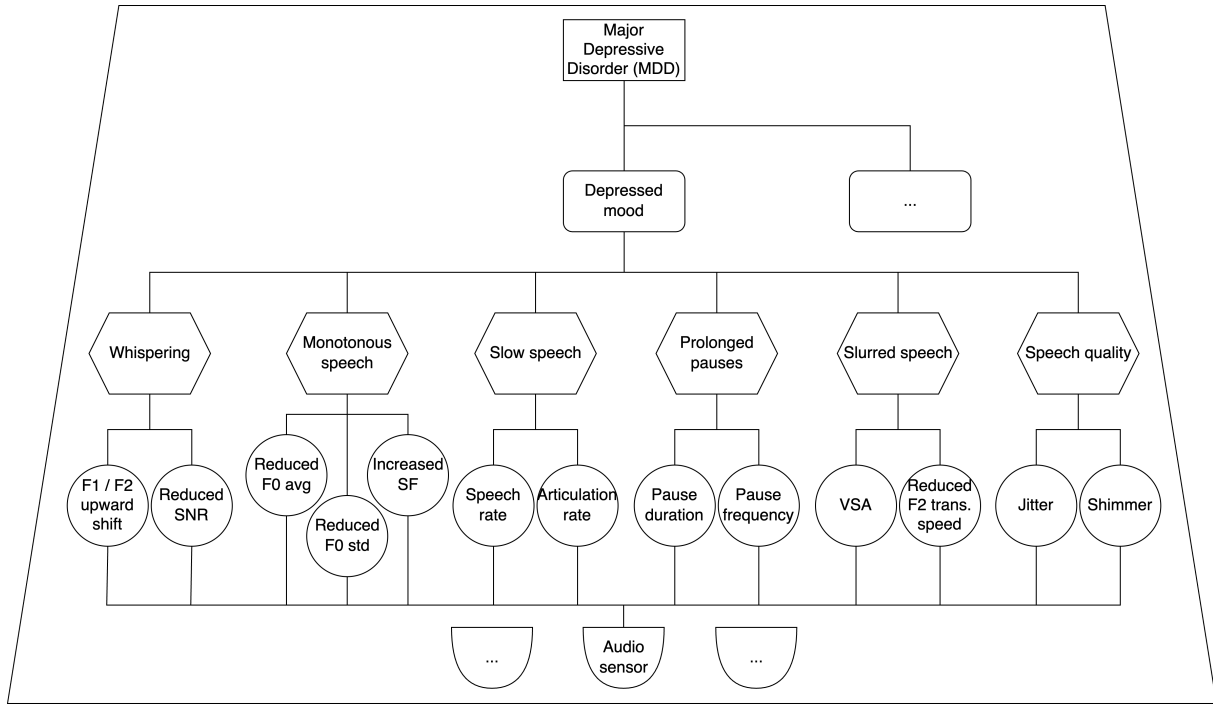


Figure 3.4: Indicator (1) *Depressed mood* mapping.

The slow speech biomarker is typically measured by the speech rate (*speech_rate* – words per time unit, including silent pauses) and the articulation rate (*articulation_rate* – words per time unit, excluding silent pauses) [130]. Closely related to speech tempo is the biomarker prolonged pauses, which is determined by the features pause duration (*pause_duration*) and the pause frequency (*pause_frequency*) as defined by Tjaden *et al.*. Pause duration refers to the total silent pause time, while pause frequency refers to the total number of pauses over the course of a speech segment [117]. Furthermore, Tjaden *et al.* investigated how slurred speech (they call it speech intelligibility) can be measured and found that slurred speech has an increased $F2$ overlap between vowel sounds, resulting in more acoustic similarity between vowels. The sum of all areas on the frequency spectrum that are associated with one single distinct vowel sound is called the vowel space area (VSA [214] – note that this feature is not extracted in the prototypical implementation in Chapter 4). Furthermore, in slurred speech, a reduced $F2$ transition speed is reported [116].

Related to depressed mood, voice features like jitter (*jitter* – pitch instability) and shimmer (*shimmer* – loudness instability) are used to measure and indicate overall speech

quality. Those two metrics are the most well-studied characteristics of speech quality in the context of mood disorders [108]. In general, spectral features such as formants and frequencies, as well as voice quality features such as jitter and shimmer, are considered the most relevant voice features for automated depression detection [10, 202]. Furthermore, prosodic features such as fundamental frequency variability and fundamental frequency average can provide information on the severity of the underlying depression [11, 104]. Among the various vocal features analyzed, formant characteristics and power spectral density features were found to be most significant for both male and female groups. In males in particular, amplitude modulation (AM) and loudness variability emerged to be equally strong predictors of depressed mood. While several studies have highlighted fundamental frequency as a promising marker for depressive behavior [125], it appears to have limited effectiveness as a standalone feature for estimating the depression severity [47, 243].

3.4.2 Loss of Interest

Figure 3.5 illustrates the proposed mapping for the MDD clinical indicator (2) *Loss of interest*. Loss of interest (or anhedonia) is a key symptom in the DSM-5 specification [18]. It is assessed through self-reporting – primarily questionnaires or guided interviews – and is exposed to individual subjective biases [188]. Yang *et al.* specifically investigated the relationship between anhedonia and voice features in patients with depression by comparing patients with and without anhedonia [188]. Investigating different voice features, they found no significant correlation between those features and any degree of anhedonia within both study groups [188]. However, the literature before that study suggests a potential link between voice characteristics and anhedonia [204].

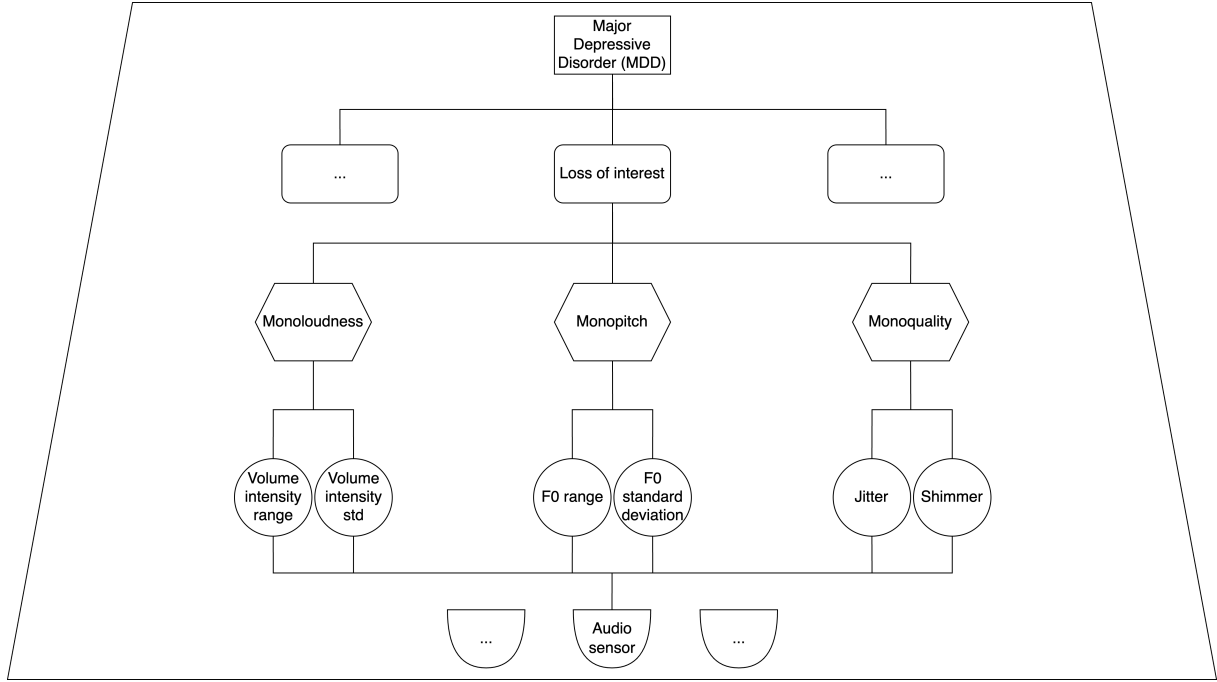


Figure 3.5: Indicator (2) *Loss of interest* mapping.

In the literature, anhedonia is described as a persistent emotional flatness and a feeling of indifference to most things [196]. During interviews with patients suffering from anhedonia, Watson *et al.* described the patient’s voices as monotonous, *i.e.*, they were talking with very little intonation and emotion [196]. However, the content of the interviews has been shown to reveal more information about possible anhedonia [196]. Subjectively, intonation and the flatness of the voice consist of monopitch (flatness in vocal pitch) and monoloudness (flatness in vocal loudness). They are captured through a wide range of different features of the prosodic voice [201]. For this work, monoquality is added to the features that describe the flatness of the voice, *i.e.*, the flatness in vocal quality metrics.

Reduced intonation (monopitch) refers to less variation in pitch over time and can be tracked by the metrics fundamental frequency range ($F0_{range}$) and fundamental frequency standard deviation ($F0_{std}$) [102]. Vocal flatness (monoloudness) refers to less variation in loudness over time and can be tracked correspondingly by the metrics volume intensity range (*volume_intensity_range*) and volume intensity standard deviation (*volume_intensity_standard_deviation*). A reduced degree of changes in vocal quality metrics (monoquality) can lastly be tracked adaptively by the metrics jitter and shimmer. Yang *et al.* found, consistent with findings from previous literature, that jitter and shimmer are positively correlated with depression severity in isolation [188]. Another study by

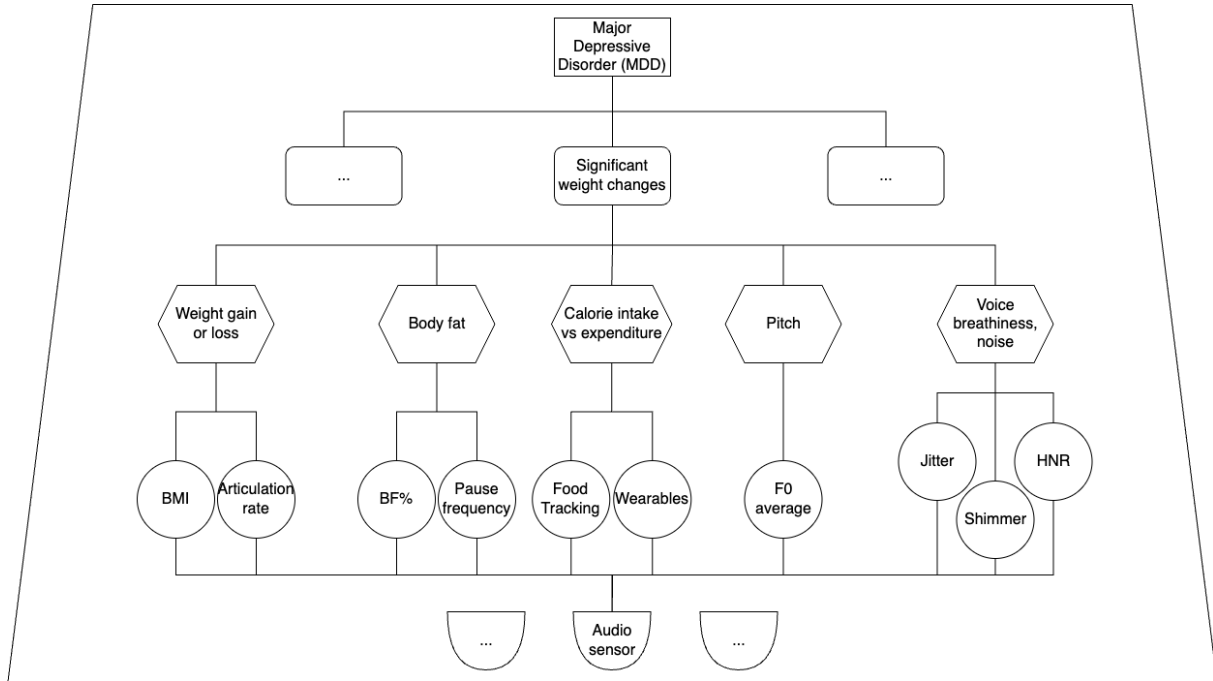


Figure 3.6: Indicator (3) *Significant weight changes* mapping.

O’Dea *et al.* employed a text analysis approach to uncover links between anhedonia and specific linguistic patterns such as the frequent use of discrepancy words [23]. Although linguistic features are not considered in this thesis, these findings are interesting, as they highlight the potential for more suitable voice features. For example, despite the evidence that acoustic voice features such as jitter and shimmer are helpful in detecting depression, linguistic features could produce better results in identifying the DSM-5 indicator (2) *Loss of interest* [202].

3.4.3 Significant Weight Changes

Figure 3.6 illustrates the proposed mapping for the MDD clinical indicator (3) *Significant weight changes*. Multiple biomarkers of diverse modalities are associated with significant weight changes: First, weight gain and weight loss are the primary biomarkers for significant weight changes. Weight cannot differentiate between body fat and other weight sources such as muscle, bone, or tissue mass. Therefore, body fat is used as a more robust biomarker [9]. Ultimately, there exist various other ways to assess weight changes, such as caloric intake and expenditure, genetics, or metabolic techniques, which are not considered in this work [65]. In fact, it is possible to detect weight changes through the effects on the voice.

A study by Redburn on how weight changes can affect the voice of professional classical singers found that both weight gains and losses substantially affect breathing, phonation, and vocal quality [210]. Obesity is associated with a decrease in vocal pitch due to slower vibrating vocal cords and breathier, noisier voice qualities [91]. However, the home setting leverages the flexibility of a wide range of different sensor types. For this indicator, the acoustic-based approach might not yield optimal results, as other sensors such as smart scales, wearables, or food trackers could be better suited to track significant weight changes in a person [97].

In this example, the aim is to illustrate how different modalities can interact within a single *Mapping Pyramid*. To assess weight gain or weight loss, the body-mass-index (*BMI*) metric is used, as this metric is a measure of weight relative to height and accounts for variations in height, age, or body composition [37]. For the body fat biomarker, the body fat (*BF%*) metric is used. The fundamental frequency average tracks the vocal pitch, and the jitter, shimmer, and harmonics-to-noise ratio (*HNR*) metrics determine the biomarker of speech quality of a more breathy and noisier voice [91].

A recent study has found that weight changes of more than 5%, both gain and loss, indicate a depressive disorder, among other characteristics such as gender, race, marital status, and various behavioral patterns [6, 8]. However, consumption of "comfort food" (often food with a high calorie content [219]) and emotional eating patterns appears to be more insightful, since foods high in fat and sugar directly stimulate the dopamine center in the brain [8]. Although a smart scale cannot capture comfort food consumption or emotional eating behavior, complementary sensors can collect such insights. Rachakonda *et al.* proposed an IoT-based smart system to monitor behavior patterns in eating stress. They offer a portable or non-portable method of tracking food intake and, ultimately, are able to detect user eating behavior and estimated daily calorie intake with an accuracy of 97% [124].

Koster *et al.* discovered in their study a reciprocal relationship between depression and significant weight changes: A depressed mood predicts a significant weight change over three years (p-value = 0.02), while inversely, people who gained or lost significant weight over a three year period had a significantly higher risk of depression (p-value = 0.03) [6]. As this thesis focuses on (potential early) depression detection, both directions of this reciprocal relationship are of interest and should be evaluated equally.

3.4.4 Insomnia / Hypersomnia

Figure 3.7 illustrates the proposed mapping for the MDD clinical indicator (4) *Insomnia / hypersomnia*.

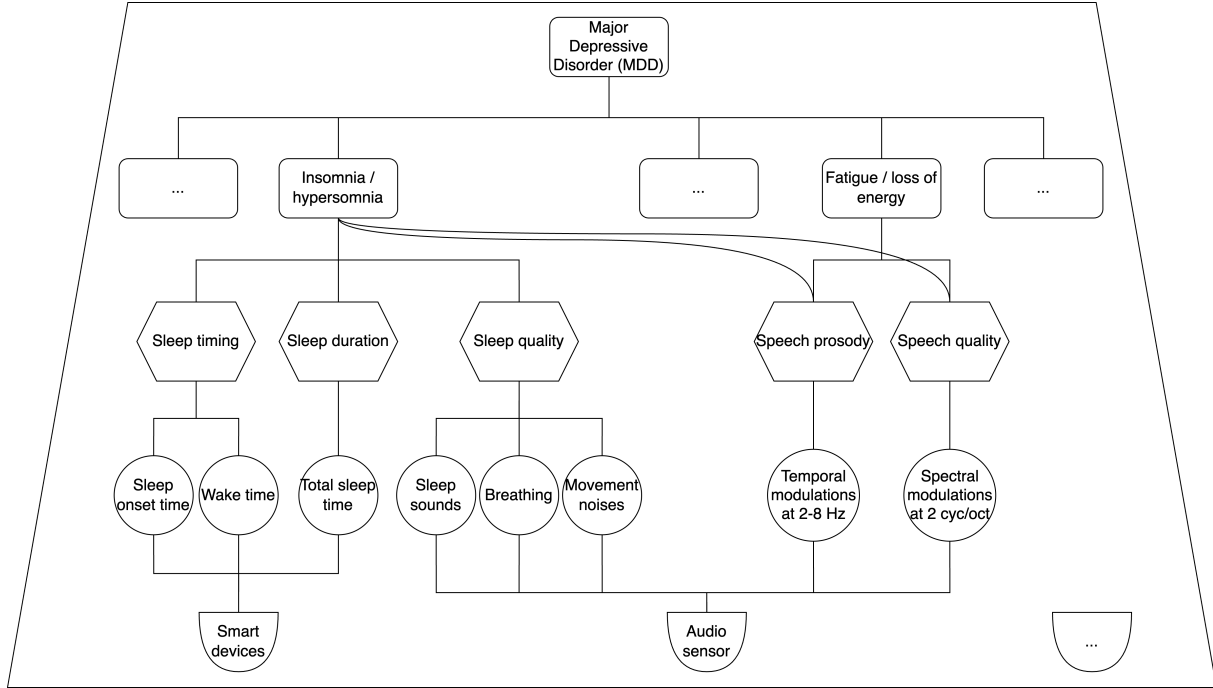


Figure 3.7: Indicator (4) *Insomnia / hypersomnia* mapping.

Insomnia and hypersomnia themselves are defined in DSM-5 and are described as the most common sleep-wake disorders in which patients suffer from quality, timing, and amount of sleep [18]. To detect insomnia or hypersomnia, it is essential to understand its diagnostic criteria, see Table 3.1, which must occur at least three times a week for at least three months [18].

Table 3.1: DSM-5 diagnostic criteria for the insomnia disorder [18].

Nr.	Diagnostic criteria
1.	Dissatisfaction with the quantity or quality of sleep.
2.	Sleep disturbances affect important areas of functioning.
3.	Insomnia is not explained by another sleep-wake disorder.
4.	Insomnia is not explained by coexisting mental conditions.
5.	Insomnia is not attributable to the effects of a substance.

The detection of insomnia and hypersomnia is its own field of research [142, 212, 223]. The biomarkers associated with insomnia and hypersomnia are mostly congruent with

the biomarkers associated with the DSM-5 indicator (6) *Fatigue / loss of energy* [191], as among the symptoms of insomnia and hypersomnia are fatigue and loss of energy itself [16, 206, 209]. Therefore, it is assumed that all biomarkers and features mapped for (6) *Fatigue / loss of energy* also apply for insomnia / hypersomnia. In addition, environmental biomarkers that directly refer to the definition of insomnia and hypersomnia are the most prevalent diagnosis criteria, including sleep timing, sleep duration, and sleep quality [18]. For the mapping of the biomarkers for indicator (6) *Fatigue / loss of energy*, see Section 3.4.6.

There is a great deal of research on different approaches, techniques, and modalities for the detection of insomnia or hypersomnia [212]. Many researchers developed sleep monitoring systems that combine multiple approaches, techniques, and modalities. The three most common techniques are home sleep apnea testing (HSAT) [123, 222], peripheral arterial tonometry (PAT) [222], and polysomnography (PSG) [123, 154, 227]. For example, Shahrum *et al.* proposed a sleep monitoring system to track user movement, sleep duration, and frequency of awakening based on the combination of camera sensors and the HSAT technique, achieving an accuracy of 90.32%, 91.80%, and 91.80% in detecting restless legs, parasomnia, and insomnia syndrome [152]. Although this thesis focuses on acoustic voice features, these findings support the general idea of an IoT-centered approach for the proposed *System Architecture*, see Chapter 4, utilizing different sensors and input modalities.

It is possible to employ an audio-centered approach for tracking biomarkers like sleep timing, sleep duration, and sleep quality. While previous research in determining the quality of sleep focused widely on PSG studies [86], utilizing the analysis of electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), and other signals [86], a recent study by Dafna *et al.* achieved an agreement with PSG of 87% by employing a ML-based approach with recorded audio signals of 30-second segments, using non-contact microphones, validating their hypothesis that audio features like sleep sounds, breathing, and movement noises capture significant information for a reliable assessment of the overall sleep quality [53]. However, it seems logical that the use of sensor data that capture metrics that are directly related to the respective biomarkers would yield the best results (also applies to indicators (3) *Significant weight changes* and (7) *Worthlessness / guilt*).

The sleep timing biomarker refers to the actual timestamp of when an individual falls asleep (sleep onset time) and wakes up (wake time) and contrasts the bedtime (lights-off

time) metric [139]. Smart devices such as wrist-worn actigraphy sensors detect these metrics by analyzing accelerometer data and heart rate drops or heart rate variability [139, 145]. To track sleep duration, similarly, accelerometer data and heart rate data can be used. The metric total sleep time (TST) is used, which is the total time an individual is asleep, excluding awake periods between sleep onset time and wake time [139].

3.4.5 Psychomotor Retardation / Agitation

Figure 3.8 illustrates the proposed mapping for the MDD clinical indicator (5) *Psychomotor retardation / agitation*. Psychomotor retardation (PMR) and agitation (PMA) are recognized as one of the most fundamental indicators of MDD. Observable symptoms include speech disturbances, facial expressions, fine motor behavior, gross locomotor activity, or ideation [44]. Bennabi *et al.* conducted a review of the status of current research on PMR and found that speech fluency, phonation, and prosody are the most researched voice biomarkers [44]. PMR occurs due to difficulties in brain-body communication that increase response time and reduce muscle activity in the body [167]. Speech fluency is characterized by speech production speed and speech interruption, while phonation is divided into speech initiation and vocal tract constriction. The prosody characteristic is grouped into the biomarkers of liveliness, articulation precision, and speech quality. In regard to the biomarkers determined for the DSM-5 indicator (1) *Depressed mood*, the biomarkers for PMR are similar, see Section 3.4.1. Ultimately, PMA is negatively correlated with PMR [1].

The assessment of speech fluency in this master's thesis is measured similarly to the biomarkers of slow speech and prolonged pauses for the DSM-5 indicator (1) *Depressed mood* through the pause duration between words [44, 57], the pause frequency, the speech rate, and the articulation rate. For the phonation biomarker, patients suffering from PMR show significant prolonged voice onset times (*VOT*), which is the time between the release of a stop consonant and the start of vocal cord vibration, and a decreased F2-formant transition speed (*F2_transition_speed*) [200], supporting the hypothesis that nigrostriatal dysfunction, such as a constricted vocal tract, can lead to changes in speech phonation [4]. Inversely, the metrics are evaluated for PMA: In contrast to PMR, PMA shows faster voice onset times and an increased F2-formant transition speed. The prosody biomarker is tracked by metrics that characterize the liveliness, articulation precision, or quality of the voice, see mapping in Section 3.4.1.

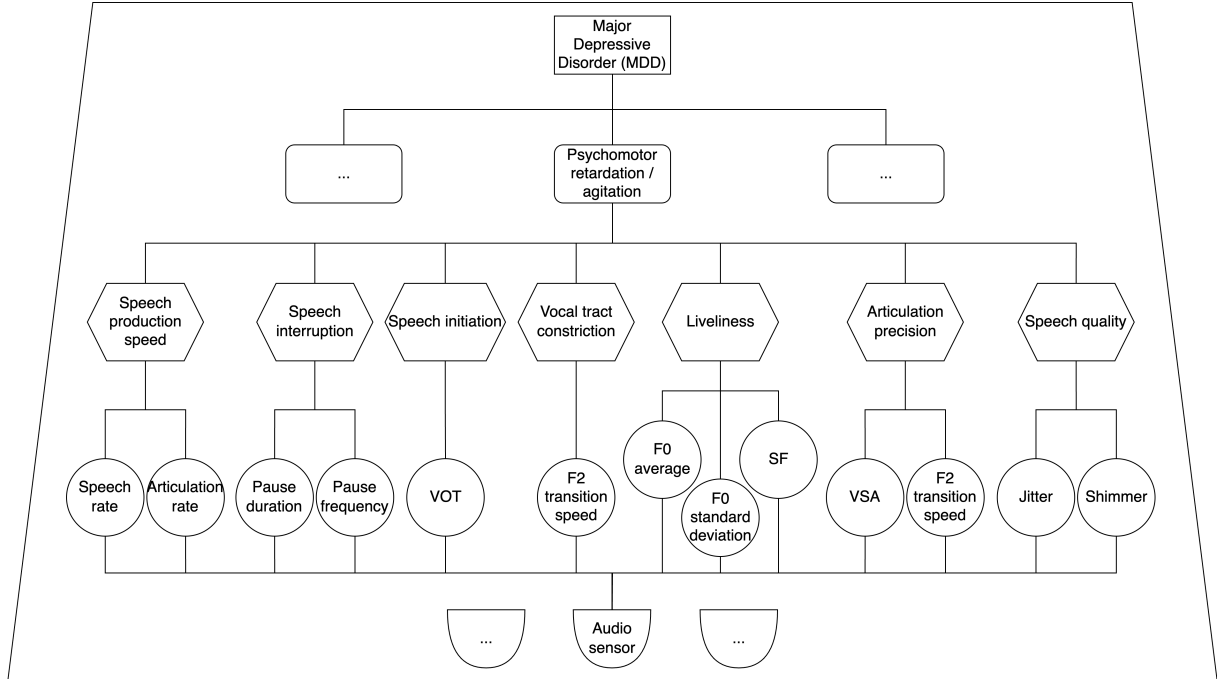


Figure 3.8: Indicator (5) *Psychomotor retardation / agitation* mapping.

3.4.6 Fatigue / Loss of Energy

Heydarifard and Krasikova found no discriminative voice features to detect insomnia (fatigue) in general [249], however, results from the work by Thoret *et al.* show that effects of sleep deprivation can be observed in acoustic voice features [58], see Figure 3.9.

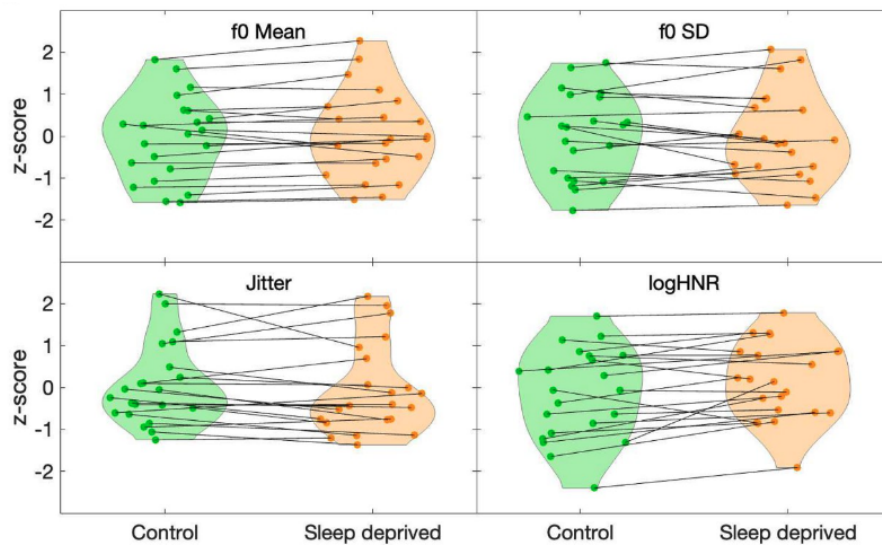


Figure 3.9: Voice features between control and sleep deprived groups [58].

Figure 3.10 illustrates the proposed mapping for the MDD clinical indicator (6) *Fatigue / loss of energy*. Fatigue can be detected by analyzing changes in a person's voice over time. The study by Thoret *et al.* shows that a lower (flattening) or a higher (animated) voice character could indicate fatigue [58]. They identified two key effects of sleep deprivation: changes in slow temporal modulations related to voice prosody and changes in spectral features that influence voice quality [58]. Therefore, the two biomarkers speech prosody and voice quality are used in the proposed mapping.

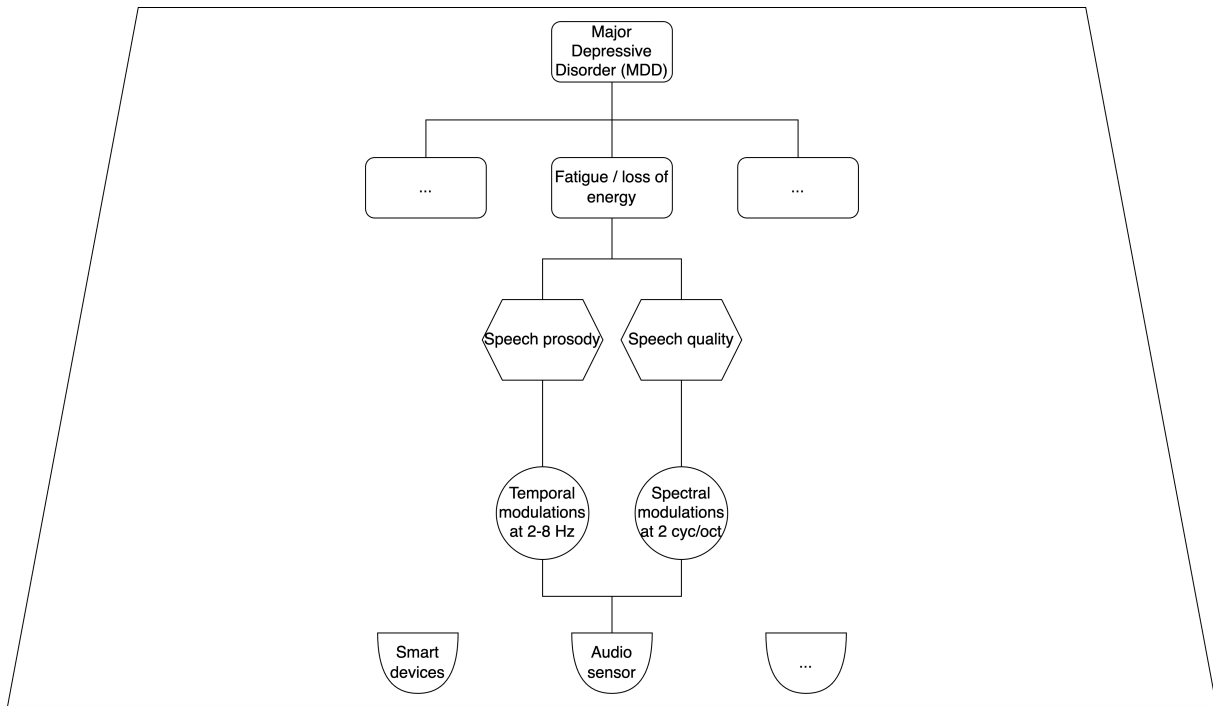


Figure 3.10: Indicator (6) *Fatigue / loss of energy* mapping.

Spectro-temporal modulation (STM) is a pre-computed representation of acoustic features. It contains the dimensions of time, frequency (spectral content), rate, and scale [58]. The frequency dimension in Hz is obtained by splitting the signal into sub-bands. Based on these sub-bands, the rate and scale dimensions are computed [58]. The rate in Hz reflects the sound amplitude modulations over time, while the scale dimension in cycles per octave reflects changes in the spectral domain over time. When analyzing the speech's rhythm and prosody, the rate dimension with temporal modulations between 2-8 Hz needs to be evaluated [58]. Similarly, when analyzing the speech's voice quality or timbre, the scale dimension with spectral modulation at around 2 cycles per octave needs to be evaluated [58]. Thoret *et al.* found that the differences in the deprived and

non-deprived groups were maximum at 3% and 11% at the population level; see the tiny differences in Figure 3.11.

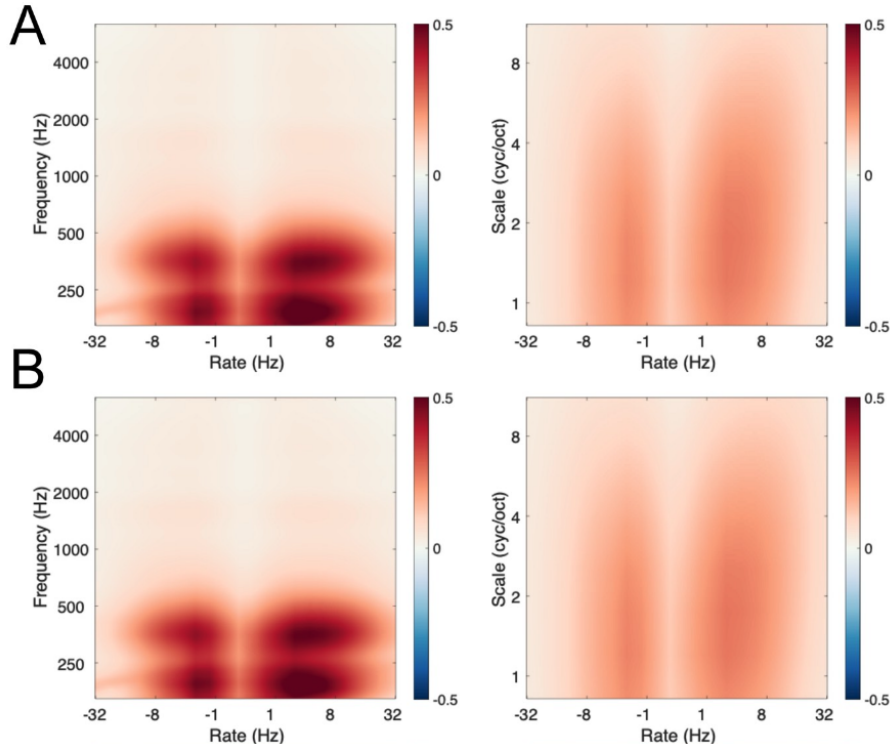


Figure 3.11: STM projections between control (A) and sleep deprived (B) groups [58].

On the individual level, on the other hand, the differences maxed around 24% and up to 40% [58]. Therefore, it is essential to perform individual-based speech analysis, as the measurements vary widely between participants [58]. As patients with insomnia have very similar voice characteristics to patients with MDD, the isolation between the two disorders becomes important. Lin *et al.* demonstrate how it is possible to reliably separate these two groups [191].

3.4.7 Worthlessness / Guilt

Figure 3.12 illustrates the proposed mapping for the MDD clinical indicator (7) *Worthlessness / guilt*. Note that the features outside the domain of acoustic voice features are drawn in the figure to underline the multimodal idea.

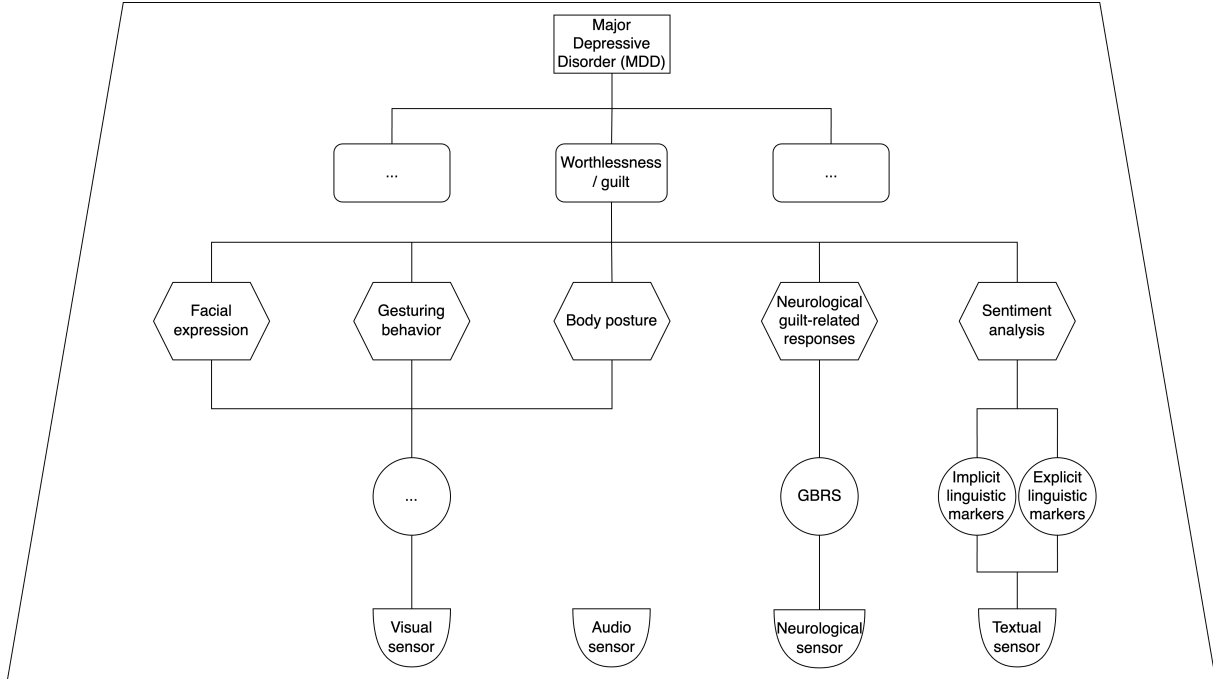


Figure 3.12: Indicator (7) *Worthlessness / guilt* mapping.

In a study by Keltner *et al.*, participants were asked to label emotions from vocal sounds that exhibit certain emotions such as guilt or shame. The results revealed that only a small proportion of the participants were able to identify guilt, many mislabeling it as sadness or reporting no emotion at all [118]. This means that those self-conscious emotions are often difficult to accurately detect through voice features [56]. Due to the current literature focusing on depression detection in a more general way, there exists an open gap of scientific findings that solely investigate biomarkers that reference and indicate the DSM-5 indicator (7) *Worthlessness / guilt*. However, there are often indirect indicators of guilty behavior in clinical contexts, even if the term itself isn't used concretely.

This highlights the broader scientific interest in uncovering hidden psychological states through observable signals in individuals [141]. Criminologists in particular have been fascinated by the identification of biomarkers of guilt emotions [141]. When the view shifts to the audio-centered approach of depression detection, the study by Gamer *et al.* from 2006 showed that voice stress analysis is not effective for the detection of guilt [141]. In addition, the integration of acoustic voice features for the detection of (7) *Worthlessness / guilt* could be detrimental to the performance and reliability of the overall system [31].

Therefore, this master's thesis proposes the use of different modalities for the detection of this indicator, leveraging the potential of the IoT environment. In this paragraph, a shortened description of potential promising biomarkers and metrics is given for this indicator. Audio data still proves to be a useful data source for the system. By transposing the acoustic signal into a textual representation, semantic and linguistic analysis becomes accessible. Meque *et al.* developed different ML architectures, including Natural Language Processing (NLP) models, focusing on the classification of guilt in textual data. They investigated linguistic markers in a nuanced approach that considers both explicit and implicit linguistic markers, achieving F1-measures between 71% and 80% [3]. In the context of neuroscience, Yu *et al.* have identified a so-called guilt-related brain signature (GRBS) that is a neurological biomarker that has been shown to be sensitive to the detection of cognitive antecedents of guilt [80]. Ultimately, among many other research directions, the nonverbal expressiveness of individuals, such as facial expression, gesture, posture, or gaze, is also being explored [31].

3.4.8 Diminished Ability to Think / Concentrate

Figure 3.13 illustrates the proposed mapping for the MDD clinical indicator (8) *Diminished ability to think / concentrate*. According to the performed literature review, there is no dedicated research that explicitly investigates how voice features can be associated with the DSM-5 indicator (8) *Diminished ability to think / concentrate*. However, when examining the diagnostic criteria for neurocognitive disorders, it becomes clear that a reduced capacity for thought and concentration is at its core [18]. According to DSM-5, neurocognitive disorders are diagnosed based on impairments in defined cognitive domains, with certain clinical thresholds that determine their severity and their subtype classifications [18]. The defined cognitive domains *complex attention* and *executive function* explicitly refer to a reduced concentration and a decreased ability to think [18]. This observation suggests that the indicator (8) *Diminished ability to think / concentrate* is, in fact, a diagnosis criterion for various cognitive disorders, which facilitates the scientific foundation for the detection of this indicator. Knowing this, the task is to find voice features that discriminate cognitively impaired individuals from the baseline and then transitively derive these voice features for the detection of the indicator. This approach helps to bridge the existing research gap for now.

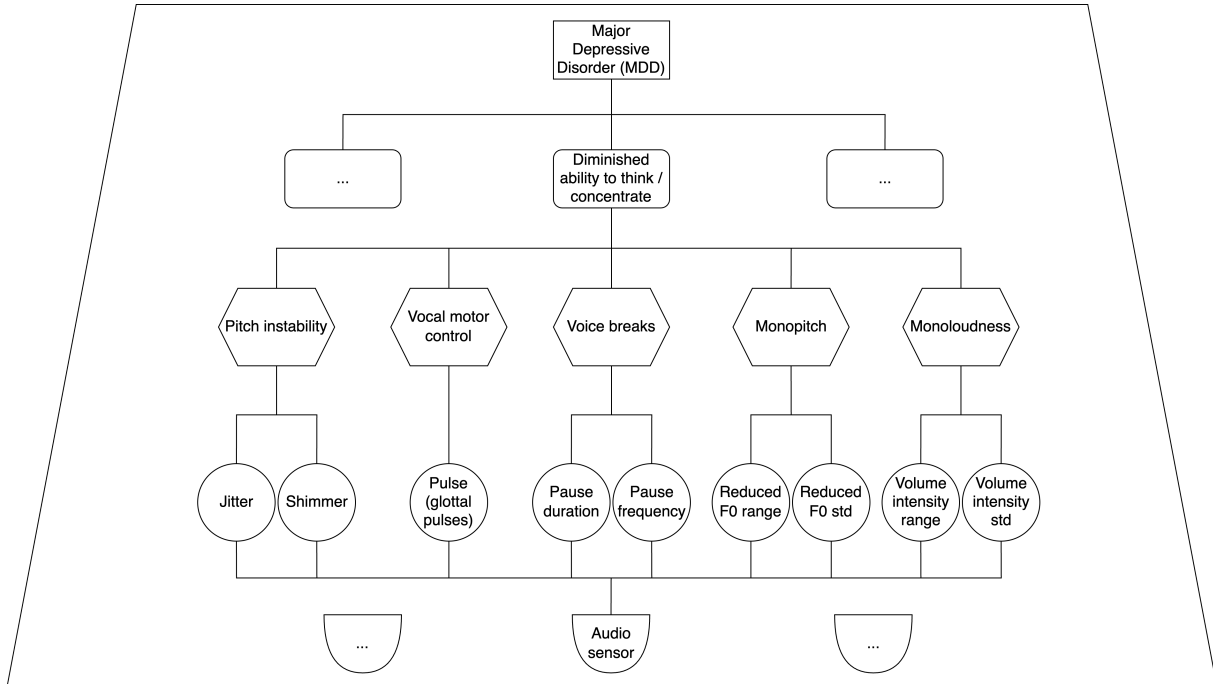


Figure 3.13: Indicator (8) *Diminished ability to think / concentrate* mapping.

A substantial body of research focuses on identifying discriminative voice features that differentiate individuals with diagnosed cognitive impairments from those without. These studies often aim to develop a model for early detection of cognitive diseases such as Alzheimer’s disease, mild cognitive impairment (MCI), and other forms of cognitive diseases [55, 98, 195, 216]. Leveraging the insights from these studies about temporal, spectral, and prosodic voice features enables ML-based approaches that already more or less reliably distinguish between normal cognition and impaired cognition [195]. Note that it is generally suggested to combine voice features with more traditional biomarkers in cognitive assessment systems [55].

Although some research approaches for the detection of cognitive disease use ML to discover patterns in voice features, they do not provide information on the detailed contribution of individual voice features to the final classification verdict [195, 216]. However, other studies provide detailed insights into the individual role of each voice feature, enhancing the explainability of the underlying mechanisms – similar to what this master’s thesis tries to achieve. For instance, Mahon *et al.* and Meilán *et al.* identified various voice features and their relevance in detecting cognitive changes and dementia risk [55, 98]. They found that cognitive difficulties are associated with increased pitch instability, decreased vocal motor control (p-value = 0.002), and a lower number of voice breaks (non-intuitively). Also, a flat and monotonous voice in terms of pitch

(monopitch) and loudness (monoloudness) (p-value = 0.002) is observed to indicate cognitive difficulties [55, 98].

The metrics jitter and shimmer are used as measurements for tracking pitch instability, while decreased vocal motor control is associated with a lower pulse (*glottal_pulses*). The biomarker of speech fluency, a lower number of voice breaks, is tracked by the pause frequency metric, indicating the percentage of total speaking time occupied by breaks or interruptions and the respective pause duration [195]. Monopitch is measured by the vibration rate of vocal folds, similar to the mappings for the indicators (1) *Depressed mood* and (2) *Loss of interest*, i.e., through a reduced fundamental frequency range, and a reduced fundamental frequency standard deviation while monoloudness is measured by the volume intensity range and volume intensity standard deviation metrics, see Section 3.4.1 and Section 3.4.2.

3.4.9 Suicidality

Figure 3.14 illustrates the proposed mapping for the MDD clinical indicator (9) *Suicidality*. Relying on vocal cues and voice features for clinical suicidality assessment is already a common practice during the diagnosis process. Clinicians subjectively interpret voice features such as verbal productivity and prosody [90]. Furthermore, studies on automated voice feature analysis for the early detection of suicide risk have shown superior accuracy compared to traditional survey-based methods [89].

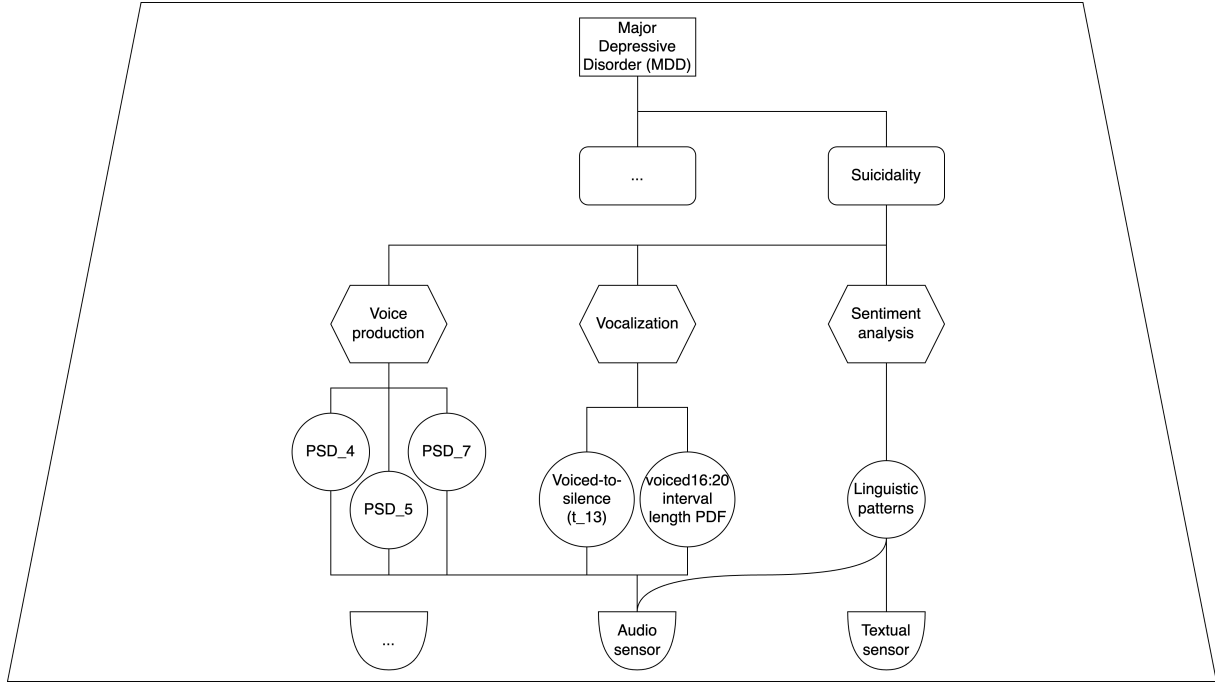


Figure 3.14: Indicator (9) *Suicidality* mapping.

There is a diverse range of biomarkers that have great potential for the early detection of suicide risk, while there is currently no consensus in the literature on the most promising biomarkers and metrics. Iyer *et al.* performed an extensive review of scientific studies on finding voice features to differentiate between high-risk suicide and baseline groups. Interestingly, the most frequently used voice features in these studies were not the most effective [89]. The highest classification accuracies were achieved using individual metrics that belong to metric groups such as speech timing patterns (median accuracy: 95%), power spectral density (PSD) sub-bands (90.3%), and cepstral coefficients of the mel-frequency (80%) [89]. Patients with a high risk of suicide often experience PMR and cognitive difficulties (noting the potential for knowledge transfer) [167]. As evaluated for indicator (5) *Psychomotor retardation / agitation*, observed effects of PMR are anomalies in speech production and speech interruption. Cognitive impairment affects the ability to clear articulation, speech interruption, and general attention [167]. Cummins *et al.* found that the voice of high-suicide-risk patients can be described by a hollow, toneless, monotonous character [162]. Homan *et al.* support this claim while adding speech interruption (particularly pause duration) and voice instability (*i.e.*, jitter) [208]. The selection of biomarkers in the proposed instantiated *Mapping Pyramid* includes vocalization, spectral features (voice production), and speech interruption.

Salomon *et al.* proposes vocalization speech features such as six transition parameters (frequency of pauses after vocalized segments in speech) and various interval length probability density functions (PDFs) that capture the duration of vocalized and silenced segments in speech [167]. They suggest that they are highly effective in identifying high-risk suicidal patients. The transition parameter *voiced-to-silence* (t_{13}) and the interval length PDF *voiced16:20* (*i.e.*, 16 to 20 frames, each 40 milliseconds of length) track vocalization in the proposed mapping [167]. These metrics capture prolonged vocalization, *i.e.*, reduced articulation switching [168].

Cummins *et al.* suggest that PSD sub-bands between 0 *Hz* and 1,000 *Hz* are the most promising candidates for further research [162]. Yingthawornsuk achieved good results using only four different PSD sub-bands. He further notes that the more PSD sub-bands are used, the better the results will be [246]. Aligning with this theory, Sanadi used eight PSD sub-bands (0 *Hz* - 2,000 *Hz*; steps of 250 *Hz*) in their work and combined the individual sub-bands to find the highest achievable accuracy. The best accuracy was achieved with PSD_4 (750 *Hz* - 1,000 *Hz*), PSD_5 (1,000 *Hz* - 1,250 *Hz*), and PSD_7 (1,500 *Hz* - 1,750 *Hz*) of 98.11% [230]. In the proposed mapping, those exact same PSD bands are used. For the biomarker speech interruption, the same metrics as for indicator (5) *Psychomotor retardation / agitation*, *i.e.*, pause duration and pause frequency, are used, see Section 3.4.5.

Note that incorporating additional modalities, such as an NLP sentiment analysis through speech-to-text translation, could further improve the detection of suicidality, as outlined by the *Mapping Pyramid* [137], see Figure 3.14.

3.5 Feature Mapping Specification

The next part of the *Linkage Framework* is the detailed specification of the selected features, *i.e.*, their associative relationship and their individual contribution to the calculation of the DSM-5 indicator scores. This specification defines how each lower-level metric is interpreted and the intermediate abstraction it represents (biomarker). Not all features contribute equally or in the same way. For each feature, the *Feature Mapping Specification* includes the following fields, see Table 3.2.

For the MDD instantiation of the *Linkage Framework*, the following *Feature Mapping Specification* has been defined, see Table 3.3. Note that for all remaining work, only this selection of acoustic voice features will be considered. Furthermore, the associative relationships and directions that are not explicitly supported by the literature as reviewed

Table 3.2: Overview *Feature Mapping Specification* fields.

Field	Description
Metric	A quantitative measure derived from raw sensor data. Represents low-level observable features.
Biomarker	An intermediate, interpretable construct inferred from the metric, representing a psychophysiological or behavioral state.
DSM-5 indicator	A clinical criterion defined in the DSM-5 that the feature contributes to.
Associative relationship	Describes how the metric influences the indicator: <i>gradual</i> (continuous effect based on value) or <i>non-linear</i> (effect has a non-linear correlation with the indicator score).
Direction	Indicates the direction of the associative relationship: <i>positive</i> , <i>negative</i> , or <i>both</i> .
Weight	A numerical value representing the relative importance or contribution strength of the metric to the indicator's overall score.

in Chapter 2 and Section 3.4 are *italic*. It is also assumed that all features are treated equally for the contributions of the individual indicator scores (weight = 1.0).

Table 3.3: MDD *Feature Mapping Specification*.

Metric	Biomarker	DSM-5 indicator	Assoc. rel.	Direction	Weight
F1 / F2 formants	whispering	(1) Depressed mood	<i>gradual</i>	positive	1.0
SNR	whispering	(1) Depressed mood	<i>gradual</i>	negative	1.0
$F0_{avg}$	Monotonous speech	(1) Depressed mood	<i>gradual</i>	negative	1.0
$F0_{std}$	Monotonous speech	(1) Depressed mood	<i>gradual</i>	negative	1.0
SF	Monotonous speech	(1) Depressed mood	<i>gradual</i>	negative	1.0
Speech rate	Slow speech	(1) Depressed mood	<i>gradual</i>	negative	1.0
Articulation rate	Slow speech	(1) Depressed mood	<i>gradual</i>	negative	1.0
Pause duration	Prolonged pauses	(1) Depressed mood	<i>gradual</i>	positive	1.0
Pause frequency	Prolonged pauses	(1) Depressed mood	<i>gradual</i>	positive	1.0
VSA	Slurred speech	(1) Depressed mood	<i>gradual</i>	negative	1.0
F2 formant transition speed	Slurred speech	(1) Depressed mood	<i>gradual</i>	negative	1.0
Jitter	Speech quality	(1) Depressed mood	<i>gradual</i>	positive	1.0
Shimmer	Speech quality	(1) Depressed mood	<i>gradual</i>	positive	1.0
Volume intensity range	Monoloudness	(2) Loss of interest	<i>gradual</i>	negative	1.0
Volume intensity std.	Monoloudness	(2) Loss of interest	<i>gradual</i>	negative	1.0
$F0_{range}$	Monopitch	(2) Loss of interest	<i>gradual</i>	negative	1.0
$F0_{std}$	Monopitch	(2) Loss of interest	<i>gradual</i>	negative	1.0
Jitter	Monoquality	(2) Loss of interest	<i>gradual</i>	positive	1.0
Shimmer	Monoquality	(2) Loss of interest	<i>gradual</i>	positive	1.0
$F0_{avg}$	Pitch	(3) Signif. weight changes	<i>gradual</i>	negative	1.0
Jitter	Voice breathiness, noise	(3) Signif. weight changes	<i>gradual</i>	positive	1.0
Shimmer	Voice breathiness, noise	(3) Signif. weight changes	<i>gradual</i>	positive	1.0
HNR	Voice breathiness, noise	(3) Signif. weight changes	<i>gradual</i>	negative	1.0
Temporal mod. at 2-8 Hz	Speech prosody	(4) Insomnia / hypersomnia	<i>gradual</i>	<i>both</i>	1.0

3.5 Feature Mapping Specification

Metric	Biomarker	DSM-5 indicator	Assoc. rel.	Direction	Weight
Spectral mod. at 2 cyc/oct	Speech quality	(4) Insomnia / hypersomnia	<i>gradual</i>	<i>both</i>	1.0
Speech rate	Speech prod. speed	(5) PSM / PSA	<i>gradual</i>	negative	1.0
Articulation rate	Speech prod. speed	(5) PSM / PSA	<i>gradual</i>	negative	1.0
Pause duration	Speech interruption	(5) PSM / PSA	<i>gradual</i>	positive	1.0
Pause frequency	Speech interruption	(5) PSM / PSA	<i>gradual</i>	positive	1.0
VOT	Speech initiation	(5) PSM / PSA	<i>gradual</i>	positive	1.0
F2 formant transition speed	Vocal tract constriction	(5) PSM / PSA	<i>gradual</i>	negative	1.0
$F0_{avg}$	Liveliness	(5) PSM / PSA	<i>gradual</i>	negative	1.0
$F0_{std}$	Liveliness	(5) PSM / PSA	<i>gradual</i>	negative	1.0
SF	Liveliness	(5) PSM / PSA	<i>gradual</i>	positive	1.0
VSA	Articulation precision	(5) PSM / PSA	<i>gradual</i>	negative	1.0
F2 formant transition speed	Articulation precision	(5) PSM / PSA	<i>gradual</i>	negative	1.0
Jitter	Speech quality	(5) PSM / PSA	<i>gradual</i>	positive	1.0
Shimmer	Speech quality	(5) PSM / PSA	<i>gradual</i>	positive	1.0
Temporal mod. at 2-8 Hz	Speech prosody	(6) Fatigue / loss of energy	<i>gradual</i>	<i>both</i>	1.0
Spectral mod. at 2 cyc/oct	Speech quality	(6) Fatigue / loss of energy	<i>gradual</i>	<i>both</i>	1.0
Jitter	Pitch instability	(8) Diminished ability ...	<i>gradual</i>	positive	1.0
Shimmer	Pitch instability	(8) Diminished ability ...	<i>gradual</i>	positive	1.0
Pulse (glottal pulses)	Vocal motor control	(8) Diminished ability ...	<i>gradual</i>	negative	1.0
Pause duration	Voice breaks	(8) Diminished ability ...	<i>gradual</i>	positive	1.0
Pause frequency	Voice breaks	(8) Diminished ability ...	<i>gradual</i>	positive	1.0
$F0_{range}$	Monopitch	(8) Diminished ability ...	<i>gradual</i>	negative	1.0
$F0_{std}$	Monopitch	(8) Diminished ability ...	<i>gradual</i>	negative	1.0
Volume intensity range	Monoloudness	(8) Diminished ability ...	<i>gradual</i>	negative	1.0
Volume intensity std.	Monoloudness	(8) Diminished ability ...	<i>gradual</i>	negative	1.0
PSD_4	Voice production	(9) Suicidality	<i>gradual</i>	<i>both</i>	1.0
PSD_5	Voice production	(9) Suicidality	<i>gradual</i>	<i>both</i>	1.0
PSD_7	Voice production	(9) Suicidality	<i>gradual</i>	<i>both</i>	1.0
t_{13} transition	Vocalization	(9) Suicidality	<i>gradual</i>	positive	1.0
voiced16:20 PDF	Vocalization	(9) Suicidality	<i>gradual</i>	negative	1.0

This *Feature Mapping Specification* is the final artifact for the complete *Linkage Framework* and concludes the work on *Obj-2*.

3.5.1 Feature Exploration

Having completed *Obj-2*, the selection of relevant features for the concrete *MDD Mapping* and the corresponding *Feature Mapping Specification*, the work focuses on *Obj-1* (*System Architecture*) and *Obj-3* (instantiation). For these objectives it is important to understand the features and their implications on the system and how they influence design decisions. Because the architecture of the system is shaped by the way features are collected, computed, stored, analyzed, and evaluated, a firm understanding of each feature is necessary. Therefore, the next step is to explore the features in more detail, which helps to define the requirements for the *System Architecture*. The following section addresses key

open questions and provides guidance for implementing the selected features effectively in a structured way.

3.5.2 Audio Descriptors

First, it is important to understand how voice features are organized. There are low-level descriptors (LLD) and high-level descriptors (HLD) [164]. Figure 3.15 illustrates the distinction between LLDs and HLDs [27].

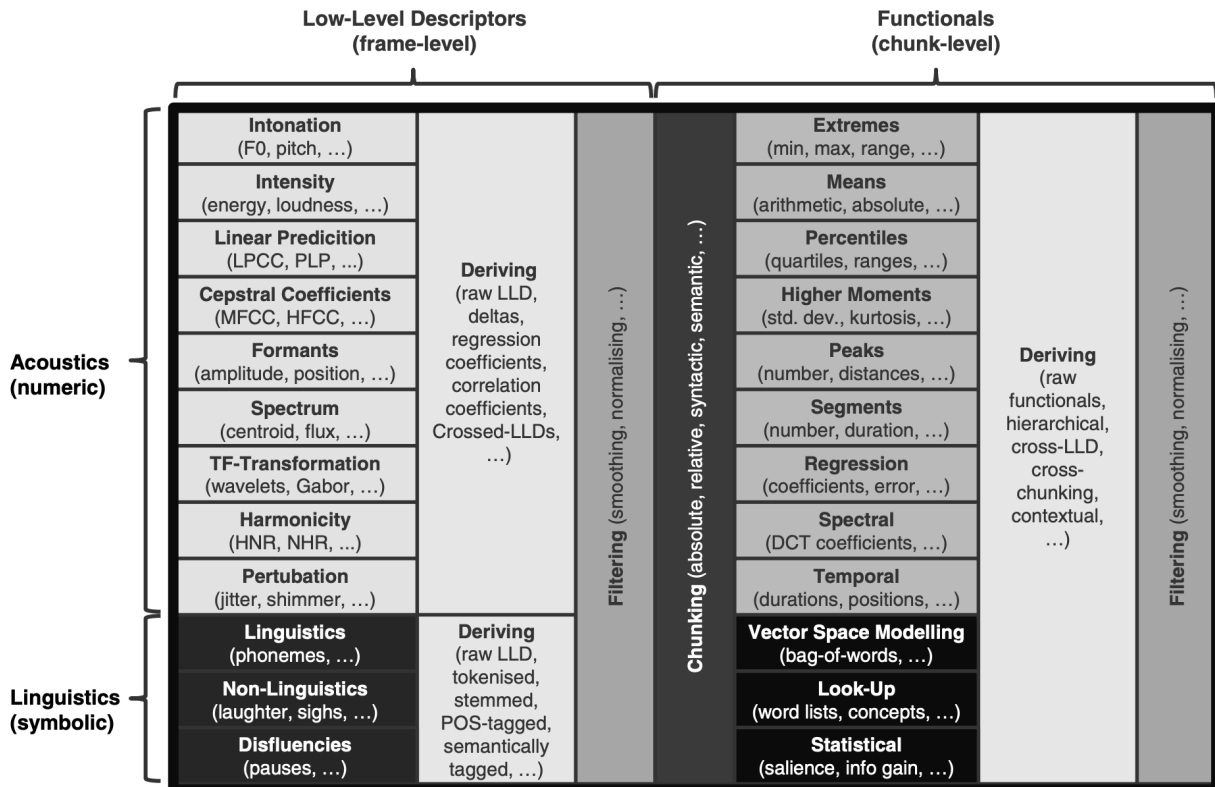


Figure 3.15: Voice metrics classification [27].

Low-level descriptors: LLDs are obtained by the raw signal itself without interpreting the sound's semantic meaning. Three granularity levels are defined for the extraction of LLDs: Instantaneous descriptors (extracted at distinct single points or very short frames); frame descriptors (aggregated over small regions); region descriptors (aggregated over longer regions) [164].

High-level descriptors: HLDs (also called *functionals*) interpret the signal using LLDs. HLDs are built by iteratively aggregating LLDs [163]. There are two types of HLDs: syntactic and semantic HLDs. Syntactic HLDs describe the structure of the sound without associating meaning to it [163]. For example, the fundamental frequency (F_0) is an LLD, where the fundamental frequency average ($F_{0_{avg}}$) is a syntactic HLD. This HLD

does not give semantic meaning to the sound. Semantic HLDs are one abstraction layer above the syntactic HLDs and refer to meaningful labeling of sounds using commonly accepted terms. They are built by a top-down classification process based on LLDs and HLDs. See in Figure 3.16 an example classification process for a *Violin* sound [163].

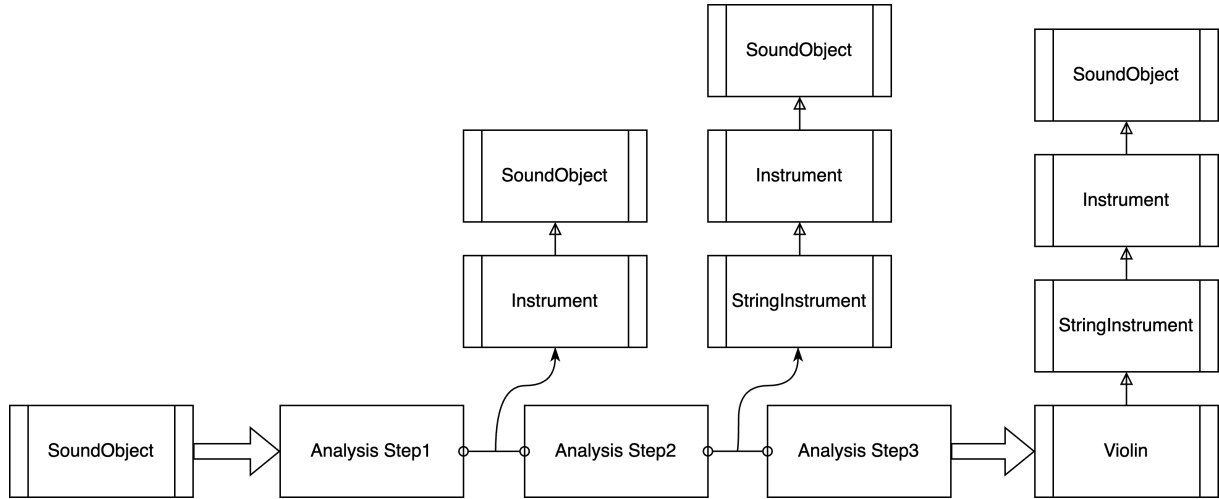


Figure 3.16: Semantic classification process for a *Violin* sound (HLD) [163].

This distinction between LLDs and HLDs in the existing literature provides a conceptual foundation. It is important to note that this classification exists on a continuum, and features may not always fit strictly into either the LLD or HLD category [165]. This grouping has important implications for the proposed system: LLDs require frame-level preprocessing. The calculation of LLDs generates large data volumes, whereas HLDs depend on individual LLDs and are not able to capture the temporal information of the audio signal [231]. In this work, the voice features are grouped into LLDs and HLDs based on whether they are instantaneously extracted from audio frames or derived through statistical aggregate functions over all frames within the audio region [62], see Table 3.4.

Various studies have proposed systematic categorizations of acoustic voice features based on their signal properties [61]. Eyben *et al.* identified four different groups of voice features: frequency-related parameters, energy/amplitude-related parameters, spectral (balance) parameters, and temporal features [61]. Similarly to that, Schuller and Batliner offer a framework for the taxonomy of acoustic features used in computational paralinguistics [27], see Figure 3.15. In this work, the categories of Low *et al.* are applied: source features, spectral features, prosodic features, and formant features [50], as shown in Table 3.4. This structured grouping brings technical specifications to light, which turned out to be useful in the actual implementation of the system. Features of the same

signal property group often need identical preprocessing and similar computation steps. For the frequency-related features, robust voice activity detection (VAD) is necessary to filter noisy sections and achieve reliable metrics [205]. Normalization and smoothing across audio samples is crucial for the energy-/amplitude-related features, as fluctuations could be introduced by different distances between the speaker and the audio sensor [157]. Spectral features are typically derived from the short-time Fourier transform (STFT). They therefore share this computationally rather expensive preprocessing step [184].

Table 3.4: Voice features descriptor type classification.

Metric	Descriptor type	Signal property group
F1 / F2 formants	LLD	formant feature
SNR	LLD	source feature
$F0_{avg}$	HLD	prosodic feature
$F0_{std}$	HLD	prosodic feature
SF	LLD	spectral feature
Speech rate	HLD	prosodic feature
Articulation rate	HLD	prosodic feature
Pause duration	HLD	prosodic feature
Pause frequency	HLD	prosodic feature
VSA	LLD	formant feature
F2 formant transition speed	LLD	formant features
Jitter	LLD	source feature
Shimmer	LLD	source feature
Volume intensity range	HLD	prosodic feature
Volume intensity std.	HLD	prosodic feature
$F0_{range}$	HLD	prosodic feature
$F0_{std}$	HLD	prosodic feature
HNR	LLD	source feature
Temporal mod. at 2-8 Hz	HLD	prosodic feature
Spectral mod. at 2 cyc/oct	HLD	spectral feature
VOT	LLD	prosodic feature
Pulse (glottal pulses)	LLD	source feature
PSD_4	LLD	prosodic feature
PSD_5	LLD	prosodic feature
PSD_7	LLD	prosodic feature
t_{13} transition	HLD	prosodic feature
voiced16:20 PDF	HLD	prosodic feature

3.5.3 Shared Information

As described in Section 3.4, there exist substantial overlaps between parts of the concrete *MDD Mapping*. To obtain an overview, the selected voice features are grouped by the associated DSM-5 indicators.

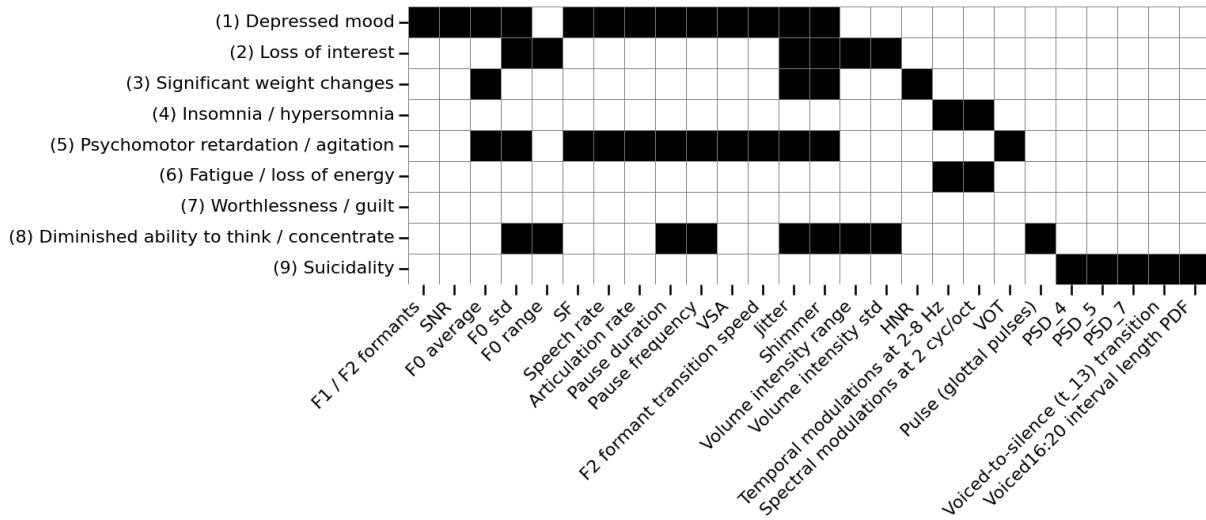


Figure 3.17: *MDD Mapping* feature intersections – matrix plot.

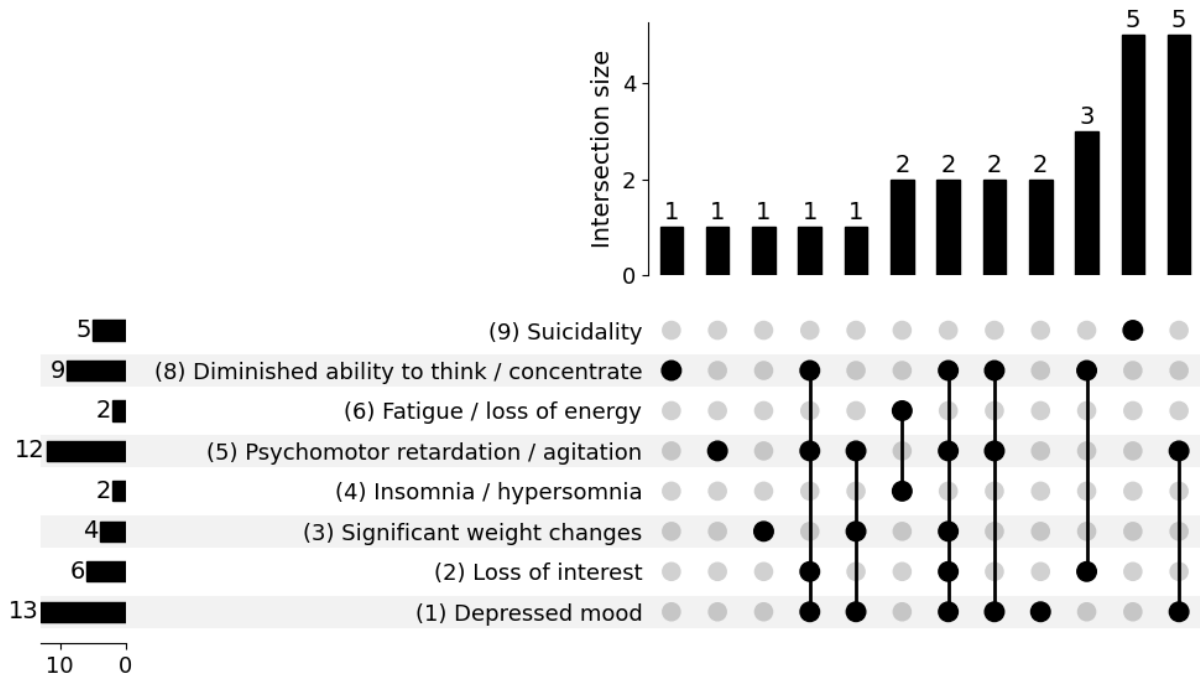


Figure 3.18: *MDD Mapping* feature intersections – upset plot.

As Figure 3.17 and Figure 3.18 illustrate, there are several intersections between voice features and indicators, which means that individual features can be used multiple

times while they are still computed only once, thereby enabling transfer of information across features and supporting a more efficient system.

3.5.4 Extraction

Given the large variety of acoustic voice features, several standardized feature sets have been proposed in the literature. Among the most frequently used feature sets are the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [61] and the ComParE 2016 feature set that was developed as part of the INTERSPEECH Computational Paralinguistic Challenge (ComParE2016) [24]. Both sets provide a range of frequently used acoustic features that span over various LLDs and the respective functionals for the HLDs. These feature sets help ensure reproducibility, improve effectiveness, and make benchmarking easier across tasks [24]. However, no feature set captures all the voice features relevant for this master’s thesis. Therefore, multiple sets of features must be used. Feature sets can be loaded into common feature extraction toolkits, such as *OpenSMILE* [62], via configuration files, e.g., *IS13 ComParE.conf*, published by *OpenSMILE* [27].

There are many open-source and research-grade audio feature extraction toolkits that facilitate the extraction of LLDs and HLDs. *OpenSMILE* (Speech and Music Interpretation by Large-space Extraction) is a feature extraction toolkit for both speech and music processing [62]. It supports both online and offline processing, enabling real-time processing and batch extraction of LLDs. Functionals (statistical, polynomial, regression coefficients, and transformations) can be applied to those LLDs [62]. *OpenSMILE* includes a built-in voice activity detection mechanism that segments audio data based on speech presence, which is very useful for the use case of continuous sensing, where the majority of audio data would be silence [62]. Note that in the implementation of this master’s thesis, a different solution is employed to address this issue. *OpenSMILE* is particularly interesting for the purpose of this master’s thesis due to its simple, scriptable, and console-based approach. Its easy configuration via configuration files and the performance-focused modular C++ architecture make *openSMILE* suitable for integrating custom components [62]. LLDs and combinations of LLDs with specific functionals already cover a wide range of the selected acoustic voice features, see Section 3.4. Two other utilized toolkits for feature extraction are the Python library *Librosa* [21] and another toolkit called *Praat(-parselmouth)* [177].

3.5.5 Analysis Approach

This final section offers important considerations regarding the extraction and analysis of voice features, as these factors impact the precision and performance of the system. This includes the methodology for the feature extraction as well as the approach to modeling the associative relationship between voice features and DSM-5 indicators.

Extraction methodology: In this work, the acoustic voice features are also classified by their extraction methodology. Most features are computed using deterministic rule-based signal processing algorithms. Others are approximations (*e.g.*, voice onset time (VOT), as the fundamental information of the voice onset is an approximation or trained itself) [237] or based on ML models trained on data to learn to compute more complex metrics (*e.g.*, facial recognition) [67]. This concept reflects whether the feature is the result of a direct mathematical operation on the signal (guaranteed precision) or an approximation learned from data (varying precision). Regarding the implementation, this categorization influences the design decisions.

Associative relationship: Another perspective is the associative relationship between the voice features and the DSM-5 indicators. As documented in Section 3.5, for this work, the associative relationship can be either *gradual* or *non-linear*. A gradual relationship with a given direction *positive*, *negative*, or *both*, simply indicates that a shift of the voice feature influences the associated DSM-5 indicator in terms of the defined direction. On the other hand, an associative relationship might be *non-linear*. This means that the correlation is exponential, logarithmic, stair-shaped with some kind of triggering threshold effects, or even more complex. In literature, those kinds of non-linear relationships are generally uncovered by using machine learning techniques [149]. For the sake of simplicity, this master thesis assumes that every voice feature has a *gradual* associative relationship. However, please note that, in reality, this is definitely not the case and must be studied in future work. The focus of this work is rather to present the concept of integrating different associative relationships in the proposed *System Architecture* and to show how they can be easily specified through the *Linkage Framework*.

Chapter 4

Implementation

This chapter is dedicated to the proposed concept and the technical implementation part of this master’s thesis. It presents the proposed *System Architecture*, outlining the design decisions made at each step and the respective reasoning behind them, referring to *Obj-1*. In the second part, the proposed *System Architecture* is instantiated together with the concrete *MDD Mapping* pyramid as defined in Section 3.3 and Section 3.5, referring to *Obj-3*.

This chapter is structured as follows: Based on the scenario description, the requirements for the *System Architecture* are formulated, against which the final implemented prototype will be evaluated. Then follows a deep dive into the proposed *System Architecture*. The second part consists of the presentation of the instantiated *System Architecture*, where each component of the system is shown and explained in detail, and design choices, specific technologies, flow diagrams, and concrete algorithms are discussed.

4.1 Scenario Description

The goal of the proposed *System Architecture* is to design a multimodal data analysis system for automated mental health diagnostics. The core idea is to provide an architecture that can easily be translated into many different application domains, such as an IoT-enabled private household setting. The design treats the scenario of an IoT-enabled private household as a first-class citizen when it comes to design decisions and the concrete implementation. This environment in particular demands continuous data collection and enables a wide range of different sensor types for input data streams – truly leveraging the multimodal system principle.

The proposed system should be able to monitor, analyze, and predict mental health disorders in general, but in this particular scenario, the primary focus is on MDD specifically. The main input data stream is an audio sensor, focusing on the speech modality for

detecting MDD. Secondly, to benefit from the IoT setting and its vast amount of possibilities for sensor deployments [106], the system must be ready for extension: Additional input data streams such as physiological sensors (*e.g.*, heart rate monitors, sleep trackers), environmental sensors (*e.g.*, light level sensors, presence detectors), and digital (*e.g.*, screen time trackers, app usage trackers) could be meaningful contributors. This demands a highly flexible and extensible *System Architecture* [134], allowing seamless addition or removal of sensors and their respective processing services. Due to the potentially resource-restricted capabilities of the home environment, the system has to incorporate smart ways for large data processing while shifting the heavy lifting from edge devices towards the system core. Given the challenges of shared spaces in a household environment that include user detection, noise, overlapping speech data, and ambiguous data, the system must incorporate reliable preprocessing, robust profiling, and mapping mechanisms to bring the data to the desired form and shape.

Ultimately, the system must present the results of the analysis in an easily interpretable manner, enabling users to understand how the system came to the given conclusions and how they are related to the collected data. This empowers the users to reflect on their personal mental health state and supports proactive mental health care. Using near-real-time capabilities and context-aware feedback, the system facilitates early and tailored intervention measures that extend beyond traditional mental health evaluations.

4.2 Requirement Engineering

To design this *System Architecture* as described in the scenario description, it is necessary to perform detailed requirement engineering (RE). This ensures a structured design approach in which needs and constraints are captured into actionable requirements while having all stakeholder perspectives considered, such as healthcare professionals, software engineers, and scientists [138]. The following customized requirement engineering process consists of three individual steps:

1. Formulate individual user stories for the interaction between the user and the system.
2. Define and describe functional requirements.
3. Define and reason about non-functional requirements.

4.2.1 User Stories

User stories make system goals clear and well-structured while improving the communication about the system and supporting an interactive and iterative design process [39]. The users of this system are healthcare professionals, software engineers, and scientists, as well as other professionals working in the broader fields of mental health or computer science. The user stories are specifically formulated with these groups in mind, enabling them to engage meaningfully with the system’s conceptual design, adapt it to support and enhance their scientific work, and access the research without being burdened by technical details. Table 4.1 contains the defined user stories for the proposed *System Architecture*.

Table 4.1: Defined user stories.

Nr.	Description
US-1	As a user, I want to easily install the system in my home so that I can start collecting data with minimal setup effort.
US-2	As a user, I want to deploy the system to the cloud so that I can process large amounts of data beyond the limits of local devices.
US-3	As a user, I want the system to configure itself automatically so that it runs silently in the background without requiring my interaction.
US-4	As a user, I want to interact with the system through the command line and a dashboard so that I can choose the interface that best suits my needs.
US-5	As a user, I want to access my data at each stage of the analysis process so that I can reflect on my mental health status and fully understand how the system comes to conclusions.
US-5.1	As a user, I want to access the raw data that has been captured by the system.
US-5.2	As a user, I want to access the contextualized data that the system has evaluated.
US-5.3	As a user, I want to access the analyzed data that has been computed by the system.
US-5.4	As a user, I want to access the DSM-5 indicator scores that have been concluded by the system.
US-6	As a user, I want to consume multiple different data streams for the processing components so that I can discover new meaningful combinations of modalities.

Nr.	Description
US-7	As a user, I want to extend the system with new sensor modalities and processing components so that I can explore their impact on the system.
US-8	As a user, I want to view historical data for each stage of the diagnostic process so that I can track my progress and tendencies over time.
US-9	As a user, I want to mark specific data points as outliers so that I can exclude them from the analysis.
US-10	As a user, I want to completely reset the system so that I can restart data collection or diagnosis under new conditions.
US-11	As a user, I want to upload my self-reported PHQ-9 responses so that the system can improve its personalized model performance.

4.2.2 Functional Requirements

The system requirements are split into functional requirements (FR) and non-functional requirements (NFR). This approach is widely recognized in software engineering and ensures that both the core capabilities (functional requirements) and quality characteristics, (non-functional requirements), are addressed equally [85]. Especially when those particularly targeted users for this work are in mind, the distinction of functional requirements and non-functional requirements is useful. While in this case the functional requirements might not be strongly influenced by the target user group, the non-functional requirements are directly dependent on them. Therefore, both are equally important in guiding design decisions and ultimately ensuring the success of the proposed *System Architecture*. Note that the differentiation between them can sometimes become fluid [85]. Furthermore, functional requirements describe the *what*, define the way the system should behave, and are considered mandatory requirements [85]. Table 5.1 contains the defined functional requirements for the proposed *System Architecture*.

Table 4.2: Defined functional requirements.

Nr.	Description
FR-1	The system must provide an installation script for the system core that sets up all dependencies and services automatically. Setup requires (1) running an installation script on the server and (2) starting a script on each sensor device.

Nr.	Description
FR-2	At startup, the system must load default configuration settings, initialize background services, and begin collecting data.
FR-3	It must be possible for the user to modify the configuration file located in the installation directory.
FR-4	The system must provide both a REST API and a web-based dashboard for accessing core functionalities.
FR-5	The system must be accessible from any device within the same local network, with no authentication required, assuming a trusted private household environment.
FR-6	The system must provide access to raw data, contextualized data, analyzed data, and the DSM-5 indicator scores via the dashboard in an interpretable format.
FR-7.1	The system must continuously capture audio from configured microphone devices or simulated from an audio file.
FR-7.2	Audio must be captured with a minimum sampling rate of 16 <i>kHz</i> .
FR-8.1	Captured audio data must be streamed from sensor devices to the central processing unit.
FR-8.2	Audio data must be encoded in mono PCM and must include metadata such as the device and a timestamp.
FR-8.3	Data must be available to multiple processing pipelines in parallel.
FR-9.1	Preprocessing must be specifically tailored for each data modality and the respective downstream task.
FR-9.2	Audio preprocessing must at least include voice activity detection (VAD), segmentation into chunks of 1-5 seconds, voice recognition, and the rejection of noise.
FR-10.1	Metric computation must follow the specifications in the feature engineering section.
FR-10.2	Metrics must be associated with a specific user.
FR-10.3	It must be possible to aggregate multiple data streams of different modalities during the metric computation.
FR-11.1	The system must maintain rolling user recognition profiles and trigger profile recalibration in case of persistent deviations.
FR-11.2	A new user profile must be created if data cannot be matched to existing profiles.
FR-12	Computed metrics must be stored in a central database with metadata such as the user and a timestamp.

Nr.	Description
FR-13.1	Contextual metrics must be computed based on recent metric trends and aggregated daily.
FR-13.2	Contextual metrics must be stored in a central database with metadata such as the user and a date.
FR-14.1	Analyzed metrics must be computed based on daily contextual metrics.
FR-14.2	Analyzed metrics must be stored in a central database with metadata such as the user and a date.
FR-15.1	DSM-5 indicator scores must be computed based on daily analyzed metrics.
FR-15.2	DSM-5 indicator scores must be stored in a central database with metadata such as the user and a date.
FR-16	The system must support data ingestion from up to 30 sensor devices, streaming data in parallel.
FR-17	The system must support the timely processing of up to 30 concurrent data streams.
FR-18	The system must support semi-structured storage of textual records to accommodate for diverse data modalities.
FR-19	User data must be stored for up to 5 years to support meaningful historical data analysis.
FR-20	The system must provide visualizations for historical data at each stage of the data pipeline.
FR-21	The system must be packaged as a containerized platform for a potential future cloud deployment.
FR-22	The system architecture must promote the easy integration and exchange of edge devices, preprocessing, and analysis components.
FR-23.1	Users must be able to mark data points as invalid to exclude them from analysis.
FR-23.2	The system must allow resetting the data of specific users.
FR-23.3	The system must allow a full data reset of every user.
FR-24	The web-based dashboard must include an interactive PHQ-9 form. Responses are used to fine-tune the user's mental health baseline, and they must be stored with metadata such as the user and a timestamp.

4.2.3 Non-functional Requirements

A non-functional requirement (NFR) describes the *how* and defines constraints and quality characteristics of a system. In general, non-functional requirements can be relaxed and can be traded off with other software quality characteristics if necessary [85]. Note that the non-functional requirements are not purely architectural quality characteristics but rather also include domain-specific requirements that are essential for the system at hand [85]. The following non-functional requirements have been described:

Architectural flexibility (NFR-1): The *System Architecture* must be designed to support horizontal scaling and modular replacement of individual components. It should allow for the effortless integration of new sensor modalities, processing techniques, and analytical methods along the data pipeline. This flexibility ensures that the system aligns with its exploratory purpose, enabling efficient adaptation towards new research questions and future scientific experiments.

Processing efficiency (NFR-2): The system must efficiently perform preprocessing of incoming sensor data, complex metric computation, and further calculations with the data along the data pipeline. Both edge devices and the individual processing components must complete their tasks within acceptable time frames in order to avoid overwhelming potential bottlenecks of the system. Additionally, storage volume capabilities must be considered to ensure long-term scalability of the persistence layer.

Operational simplicity (NFR-3): The system must provide an easy setup and usage experience, promoting research-driven experimentation. This includes straightforward deployment strategies and simple user-facing interfaces. Particular emphasis should be put on the straightforward configuration of the conceptual mapping between metrics and DSM-5 indicators, as well as the possibility to configure personalized user baselines in a modular way with minimal engineering overhead.

Explainability (NFR-4): The system must provide transparent and interpretable outputs that allow users, researchers, or clinicians to understand how the individual sensor data metrics contribute to the DSM-5 indicator scores. Each result should be traceable through the defined mapping configuration and thereby build trust in the system's assessments.

Adaptability (NFR-5): The system must build a personalized behavioral baseline for each user. As user behavior evolves over time, the system incrementally adapts the baseline, ensuring more accurate and individualized DSM-5 indicator scores.

Practical soundness (NFR-6): The system must produce results that are practically sound and well-aligned with established diagnostic standards such as the DSM-5 and the PHQ-9 questionnaire. This ensures that computed indicators are meaningful, credible, and potentially useful in clinical assessment processes. The system's output should closely correspond to results obtained through other mental health assessment methodologies.

4.3 Architecture

4.3.1 Architectural Style

The architectural choices for the proposed *System Architecture* are driven by the requirement engineering where user stories guide the functional requirements that define the system's behavior. These, in turn, must be implemented in accordance with the quality characteristics determined by the non-functional requirements.

In software engineering, the choice of the software architecture for a given project is highly influenced by the defined non-functional requirements. Therefore, they serve as selection criteria when it comes to architectural style decisions [43, 193]. But not only the non-functional requirements are important for choosing the right architectural style. As Cervantes and Kazman suggest, also the design purpose, the primary functionalities, architectural concerns, and constraints are key drivers for the architectural design [73].

The proposed architecture follows a platform-based approach. A platform in software engineering is defined as a complete infrastructure that provides the basis for building and executing different applications [73]. The platform-based design approach supports a structured, layered development process, where the "meeting-in-the-middle" methodology allows for effective design-space exploration. By defining clear abstraction layers, designers can independently evaluate decisions in terms of performance, flexibility, and reusability [7, 14]. This methodology also promotes the encapsulation of complexity, allowing developers to focus on system behavior without being hindered by low-level implementation details. Ultimately, platform-based design enhances modularity, as each design layer is composed of well-defined and interoperable components [7, 14]. A widely adopted architectural instantiation of this approach is the Microservice Architecture (MSA), which, in combination with API-driven interfaces, forms a robust system platform stack [7].

Therefore, for the implementation of the system, an MSA approach is followed. The MSA is an architectural style for distributed software systems, which consists of fine-

grained services with their own life cycles [153]. Among the many quality characteristics are its flexible configuration, ease of development, increased modularity, and easier maintenance [153]. Generally, MSA is the natural fit for cloud development, which has previously been formulated as a user story [78]. The household environment opens many possibilities to use various data sources. In IoT, our surroundings are filled with many objects that are connected to the local network in some way. This leads to enormous amounts of data that must be collected, stored, and processed [96]. IoT devices are typically limited in resources with little storage and processing power capabilities [211]. The "cloud-centric" vision of IoT systems views the internet or cloud services as the main components, while the data is provided by the "things" (here, edge devices). Gubbi *et al.* proposed such a cloud-centric IoT interaction framework that builds a flexible and open architecture that allows interaction in a suitable manner based on the requirements at hand [96]. While the collection of the data still needs to be done locally with edge devices, this cloud-centric approach leverages the potential of virtual infrastructure of the cloud for a distributed storage and computing model [96]. This section is dedicated to the discussion of why the MSA is a reasonable fit for the proposed *System Architecture*.

Scalability: In an MSA, each component is an independently deployable unit. This allows individual services to scale their resources independently of one another, enabling efficient horizontal scaling. As a result, the architecture is inherently highly scalable. Additionally, the literature offers numerous architectural patterns specifically designed to enhance the scalability of microservices-based systems [35, 207].

Modularity: As the proposed system has clearly separable boundaries between its responsibilities, modularity can be easily achieved with the MSA. Each microservice can encapsulate a specific functionality, such as data ingestion, preprocessing, analysis, or user profiling, allowing independent deployment. This separation also supports technological heterogeneity and resilience, which are critical characteristics for research systems such as this one [228].

Extensibility: The MSA is a good choice for complex systems because of its focus on extensibility [48]. The decoupled structure is ideal for this research system, where new experimental components (*e.g.*, additional sensor types or different computation services) can be frequently added to the system. Each feature is independently developed, tested, and deployed, where following an event-driven architecture approach (which introduces greater complexities) further facilitates the concept of loosely coupled in distributed environments [171].

Usability and simplicity: Setting up a data pipeline for mental health monitoring includes many complex steps that take a lot of time to implement [74]. The proposed system provides the user with all necessary building blocks, with clearly separated responsibilities such as user profiling, metric extraction, data storage, etc. By creating individual services, the system becomes more usable, as this modular structure allows developers and users to better comprehend and modify individual parts of the system without needing to understand the details of the architecture and its implementation.

Deployability: Because of the exploratory and scientific nature of the proposed system, the user might introduce many small changes to the software to test ideas. In monolithic applications, the whole system needs to be tested and deployed to release such changes. The bigger those changes are, the more likely it is that something goes wrong. The MSA allows the user to test and deploy individual services independently, thus reducing headaches during deployment processes [171]. Modern CI/CD technologies further reduce the complexity of this process, particularly for MSA systems.

4.3.2 Design Instantiation

This section presents the proposed *System Architecture* instantiation with its related design decisions and implications. It gives an overview of the data flow and how the components interact with each other. Each individual component will be discussed in detail and referred to the respective software artifacts on GitHub, see Appendix D. When referencing artifacts on GitHub, only the filename, method name, or entity name is mentioned. Further information about the repository itself can be found in the *README.md* file [128].

4.3.2.1 High-level Design

The high-level architecture, see Figure 4.1, describes the proposed *System Architecture* design on a coarse-grained level, with its involved key components and the respective high-level design considerations, including service granularity, service-level architecture, data ownership, data access, data persistence, and service communication.

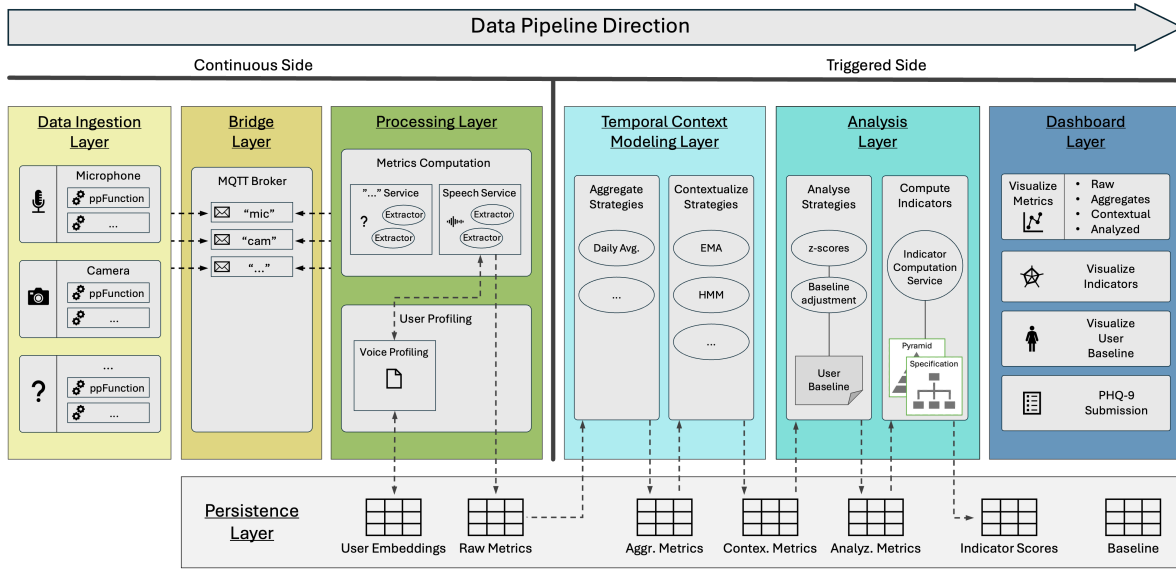


Figure 4.1: High-level architectural design – component diagram.

The design's data flow follows a left-to-right structure, reflecting the progression through the data pipeline. At each stage, dedicated services are employed with clearly defined responsibilities and constraints. While components typically communicate with their immediate neighbors, certain use cases require interactions between non-adjacent components. The system is divided into a *continuous side* and a *triggered side*. The *continuous side* processes the streamed data as it arrives in a near-real-time manner, whereas the *triggered side* only progresses through the individual pipeline steps when the user triggers the computation manually. This approach has been chosen due to potential processing bottlenecks in the rather expensive computation steps involved in the *triggered side* of the system.

Service granularity: The system follows a hybrid service partitioning approach, combining both domain and technical partitioning [148]. The domain partitioning focuses on grouping functional responsibilities (*What is a service doing?*), while the technical partitioning addresses infrastructural concerns (*How is the service deployed?*) [148]. The result of the partitioning is a set of bounded contexts (single deployable units), and ultimately implemented as microservices [148]. As illustrated in Figure 4.2, the bounded contexts encapsulate cohesive sets of unique functionalities. While some bounded contexts are grouped by their functionality (*temporal context modeling, analysis, dashboard, user profiling*), other bounded contexts are grouped by their deployment strategy (*bridge, data persistence*). The services *data ingestion* and *metric computation* are special cases. They could be grouped by their functionality, but due to the fact that

the sensing devices are mostly highly resource-constrained, they have different hardware requirements and have to be deployed independently.

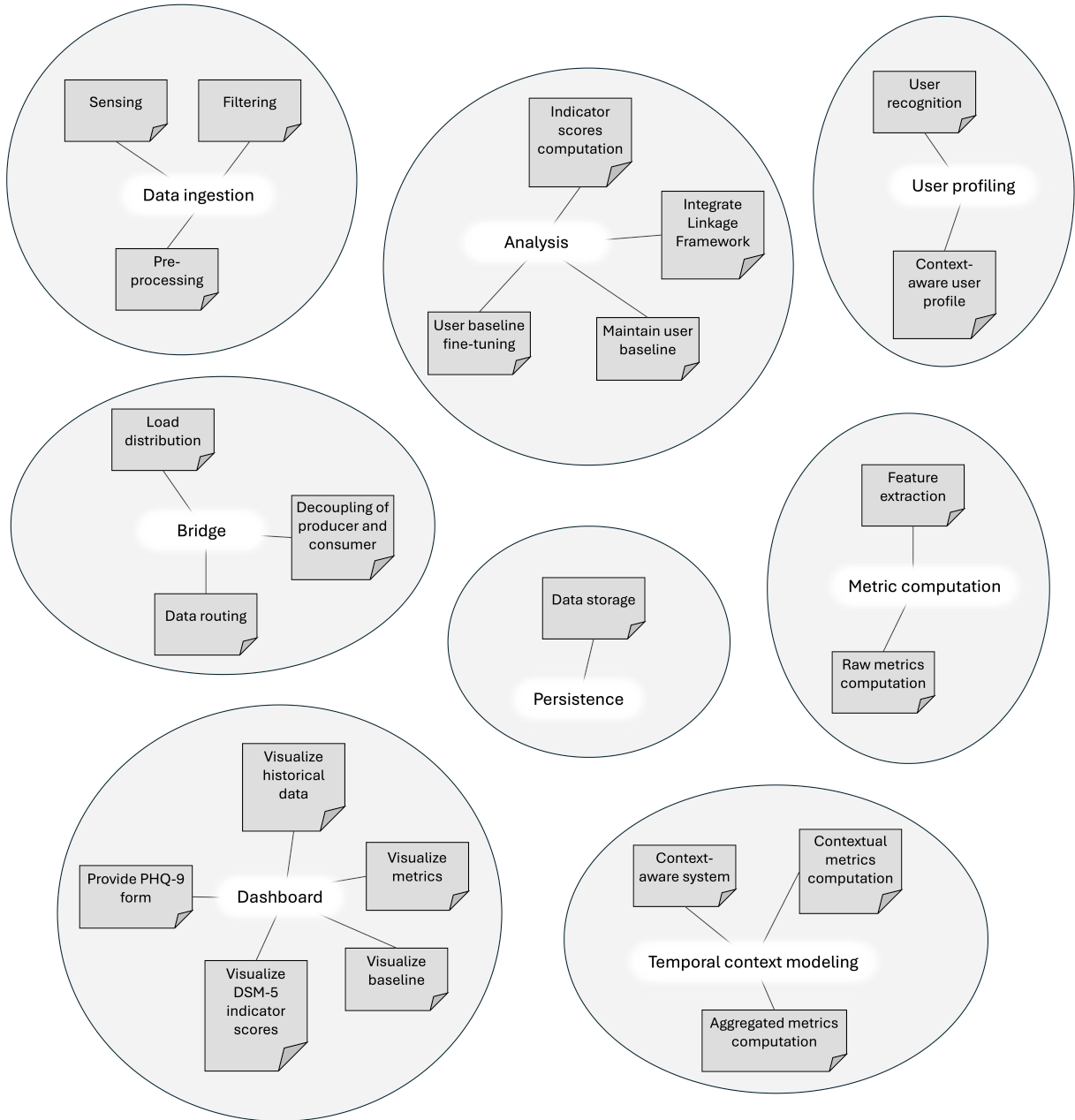


Figure 4.2: Modeled bounded contexts.

Service-level architecture: The implementation paradigm of the individual bounded contexts depends on the given requirements. In this system, components that are expected to be replaceable or extensible in the future are implemented following the "Hexagonal Architecture" pattern (also known as "Ports and Adapters") [2]. This architectural style proposes a consistent internal structure across all services, where ports define standardized contracts, and adapters provide implementations for the ports. This ensures

that the conforming services remain interchangeable and easily integrable [2]. While the architectural style defines a strict structure, the proposed implementation here adopts a more lenient interpretation for simplicity. However, the contract-based design is essential to meet the extensibility requirement as defined in Section 4.2.3. An overview of the internal structure of each component is presented in Table 4.3.

Table 4.3: Service-level architecture paradigms.

Name	Structure	Note
Data ingestion	Manual script	Collection of inter-dependent Python scripts.
Bridge	Infrastructure service	MQTT broker as Docker image by Mosquitto [187].
Metrics computation	Ports and adapters	Well-defined standardized ports and adapters for exchangeability.
User profiling	Ports and adapters	Well-defined standardized ports and adapters for exchangeability.
Data persistence	Infrastructure service	MongoDB instance as Docker image [158].
Temporal context modeling	Ports and adapters	Well-defined standardized ports and adapters for exchangeability.
Analysis	Ports and adapters	Well-defined standardized ports and adapters for exchangeability.
Dashboard	Monolithic client-server	"Stream-lit" application: Python backend and client-side rendered frontend – Both hosted on the same machine [88].

Data ownership: Data ownership refers to the notion of which service owns the data. Typically, data ownership belongs to the service that performs write operations in the database [166]. There are different types of data ownership, like "single ownership" (only one service writes to the database), "joint ownership" (multiple services write to the database), and "common ownership" (almost every service writes to the database) [166]. In this system, every service writes data to the same database, referring to common ownership. However, on the collection level, each service group writes data to its own

dedicated collection, referring to single ownership [166]. Note the indication of service group: In some of the bounded contexts, *e.g.*, *metric computation*, multiple services can be deployed in parallel to leverage the metric computation of different modalities in the future. Every service of the same service group writes to the same collection, referring to "joint ownership". Thus, a hybrid data ownership approach is utilized [166].

Data access: Similar to data ownership, the employed concept for data access has to be tailored for each service individually. Here again, multiple popular data access patterns exist. There are only four services that require reading the database: *user profiling*, *temporal context modeling*, *analysis*, and the *dashboard*. The *user profiling* service employs "in-memory caching" for accessing the voice embeddings for the users in the system [166]. The services *temporal context modeling*, *analysis*, and *dashboard* rely on no specific data access pattern, as the *triggered side* of the data pipeline only performs operations upon specific request by the user. Since performance and responsiveness are not explicitly requested by the non-functional requirements, there is no need for a more sophisticated data access pattern.

Inter-service communication: To support the concept of easy extensibility, the inter-service communication in this system is intentionally kept minimal and straightforward. This design decision facilitates the integration of new sensors, modalities, metrics computation, and analytical components by the user. As shown in Figure 4.1, an asynchronous communication model is employed between edge devices and the *metrics computation* service group through a dedicated *bridge* component relying on MQTT. This "publish-subscribe" mechanism decouples the *data ingestion* from the (heavy-lifting) downstream processing services, making it particularly easy to integrate new sensor modalities without impacting the rest of the pipeline [148]. For the communication between other services – such as initiating the *temporal context modeling* or *analysis* computations, which are triggered by user activity from the user-facing *dashboard* – a RESTful approach is employed [87].

4.3.2.2 Data Ingestion Layer

The central responsibility of the proposed *System Architecture* is to compute and track biomarker data for individual users in an IoT-enabled private household environment. Therefore, the first core objective is to model a process for ingesting raw input data. Given the diversity of input modalities (*e.g.*, audio sensors, smart devices, mobile phones, sleep trackers, physiological sensors, calendar or weather data, social media data, exter-

nal API data, or other contextual information), varying data models must be handled, and appropriate collection methods provided. Therefore, each data ingestion method must be specifically tailored to the unique characteristics of its source. This calls for a generalizable approach, one in which the data ingestion process is uniformly modeled across modalities to enhance scalability, interoperability, and maintainability. However, note that within this work, the primary input modality is audio-based voice data, and the implementation focuses on audio sensor data coming from microphones.

In modern data-driven systems, particularly in multimodal environments like IoT systems, integrating heterogeneous data presents challenges related to volume, variety, and velocity that are well-recognized and have been addressed by numerous industries and organizations [84]. Due to the high complexity of such systems, organizations often rely on the concept of a *data ingestion layer* [19]. The *data ingestion layer* is the entry point for data into a system, and large companies such as Netflix, LinkedIn, or Uber have built data ingestion pipelines using a wide range of different technologies, tailored to their specific needs of their use cases [197].

The approach of this master’s thesis for modeling a generalized and customized data ingestion strategy is inspired by data ingestion frameworks. This presents a structured approach to designing, implementing, and managing the system for heterogeneous input sources. Typically, data ingestion frameworks consist of the following core concepts [197]:

- Data collection: Collecting data from many sources with multiple formats, structured, semi-structured, or unstructured.
- Filtration process: Filter and sanitize data by parsing, removing redundancies, deleting invalid data, or handling null data values.
- Transportation process: Route the data to the respective downstream data consumers such as databases or other services.

The data ingestion of this implementation can be found in the source code on GitHub within the *data_ingestion_layer* directory, where two different folders can be found [128]:

1. *framework* – contains abstract contracts to provide common edge device behavior across generalizable sensor types [128].
2. *implementations* – contains instantiable modules that can be used, modified, or created by the user [128].

The top-level entity is the *AbstractEdgeDevice* class, as shown in Figure 4.3. All devices in the system that are collecting data from different sources inherit this base class. The class defines the contract by which all connected edge devices must conform. Furthermore, this provides a uniform way to interact with the edge devices through the defined methods *run()* and *stop()*. By this, the generalization of the three-fold process consisting of *collect*, *filter*, and *transport* is imprinted on every edge device. Depending on the modality, further specifications can be added towards this base class. Furthermore, it builds the setup for communication with the *bridge layer*, i.e., the MQTT broker. To be more specific, for example, the abstract class *AbstractAudioDevice*, inherits the *AbstractEdgeDevice* class and further defines standard behavior. It provides a default implementation for the orchestration of the described *collect-filter-transport* data flow. Furthermore, the class implements the *transport(filtered_data)* method, as every child of the *AbstractAudioDevice* class intends to publish messages to the *bridge layer* in the same format. The *framework* directory contains a directory, *payload*, for storing data classes which are another technique for standardizing messages to the *bridge layer* of the same modality.

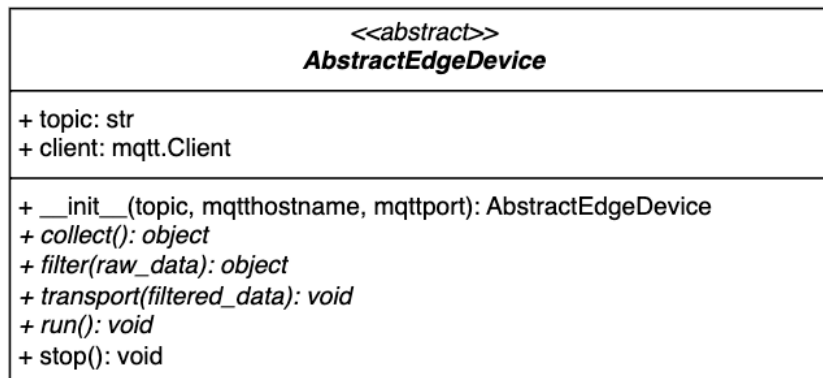


Figure 4.3: *AbstractEdgeDevice* – UML class diagram.

Figure 4.4 illustrates an excerpt from the class diagram for the *data ingestion layer*. For now, the work has implemented audio types of edge devices, such as reading the data from an audio stream or from a *wav* file. On the left side, another modality is implemented with textual data. Similarly to the audio modality, there are potentially many different data sources with the textual modality. The proposed module architecture allows for easy standardization across modalities and specification within the different modalities. Additionally, the approach with predefined payloads for the messages helps in predicting fault tolerance, as errors will become easy to track in the data pipeline.

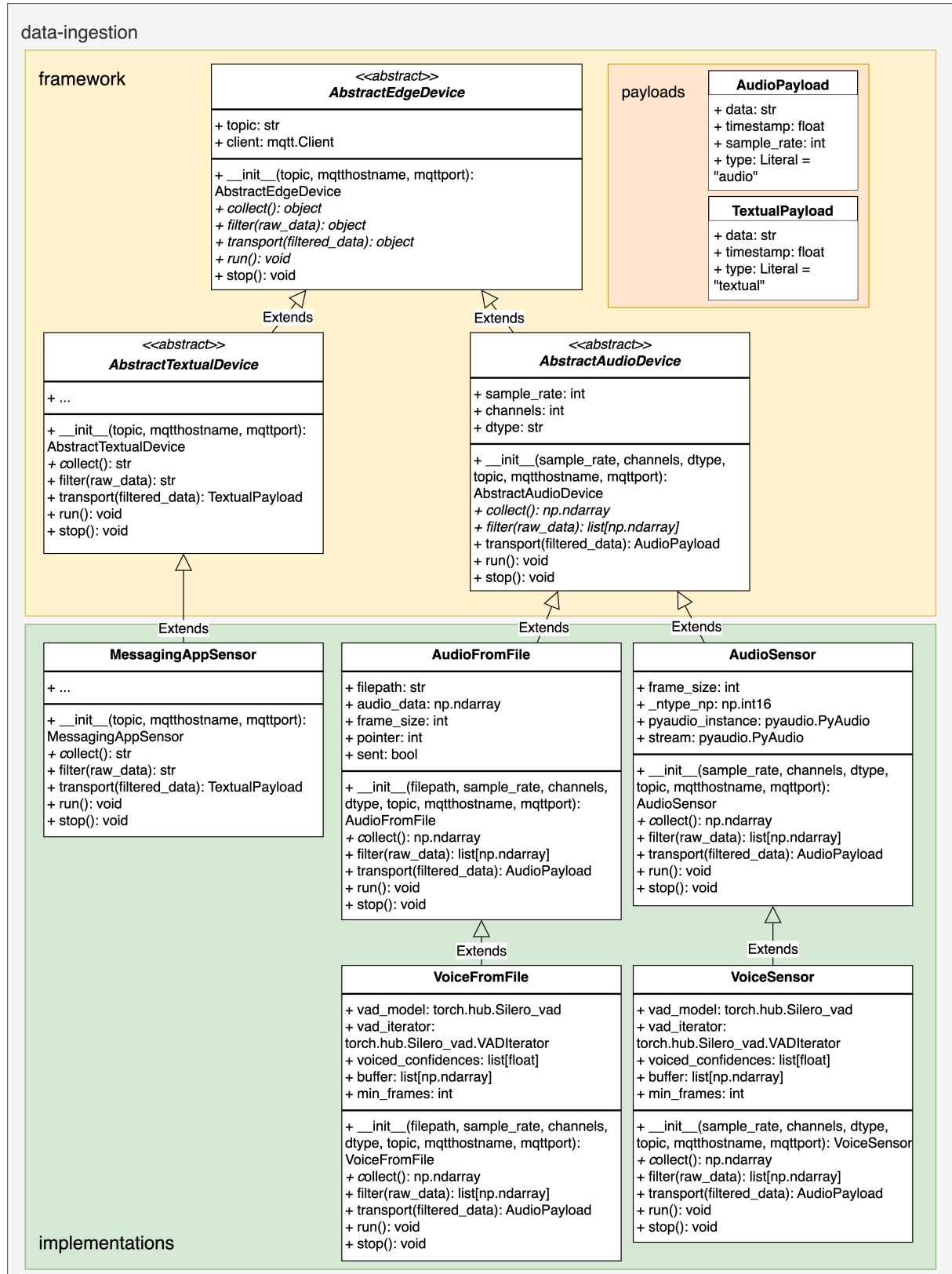


Figure 4.4: Data ingestion layer – UML class diagram.

Keeping in mind the general concept of a cloud-centric system, the *data ingestion layer* aims to employ "edge computing" to some degree in order to reduce network load and keep computing effort minimal over the system [99]. The idea is to process the data where it becomes available. This edge computing technique is typically found in the *filter* step of the *collect-filter-transport* process.

During the *filter* step, the implementations *VoiceSensor* and *VoiceFromFile* perform voice activity detection (VAD) on the perceived audio data [128]. In a typical household environment, there are often long periods of silence or background noise, with no one speaking. In these cases without filtering, the *data ingestion layer* would simply stream such useless data continuously to the *bridge layer*. Due to VAD, bandwidth and computation capabilities can be preserved and are only used in case the VAD is detecting longer periods (>30 frames, 512 samples each). This implementation relies on the open-source library *SileroVAD* [224]. *SileroVAD* is a lightweight, pre-trained, enterprise-grade voice activity detector. It processes audio chunks of 30 milliseconds in less than 1 millisecond and is trained on a multilingual corpus (6000+ languages), which generalizes well across various environments and noise conditions [224]. *SileroVAD* stands out for its reliable results on speech detection tasks, which is crucial for the proposed system, as the household setting might confront the system with many different noise scenarios. Its lightweightness enables resource-constrained edge devices to perform VAD (2 MB memory required) and *SileroVAD* has its typical use case in IoT systems [224]. As in this system, only voiced segments should be transmitted to the *bridge layer*, the input audio data stream will be filtered by *SileroVAD* with a speech prediction confidence threshold of above 0.5 to filter out noise and silence.

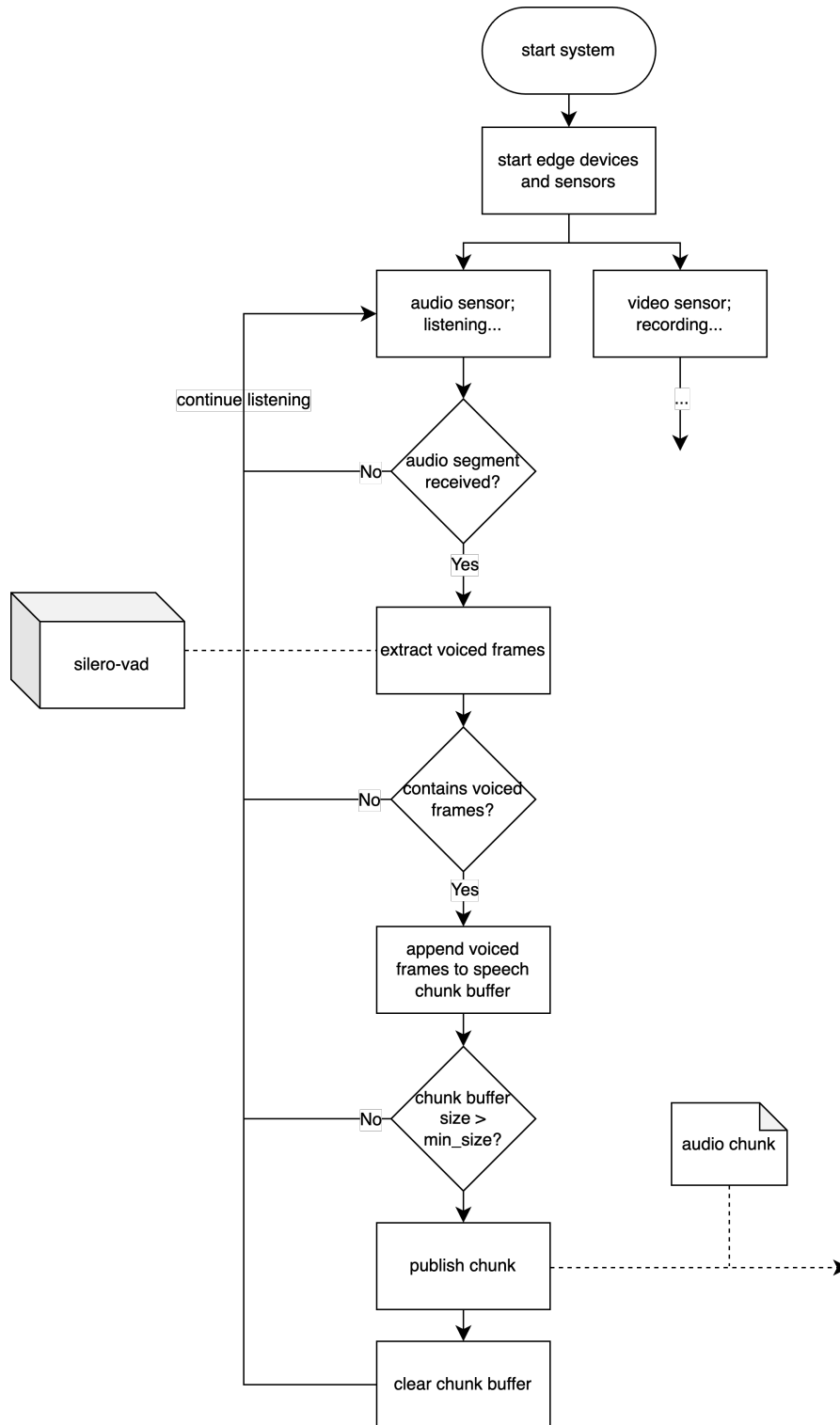


Figure 4.5: Process logic of *data ingestion layer* – flow diagram.

Data ingestion is performed through batch or stream processing and is decided on the basis of the requirements of the speed with which the data is needed for analysis [218]. In this master's thesis, the selected processing strategy is dependent on the input

data modality. For example, audio data is streamed in real-time but only forwarded to the *bridge layer* if voiced segments have been detected. Textual data – possibly coming from a real-time messaging application – could be processed in batches, reducing the communication between producer and consumer. After data *collection* and *filtration*, the *transmission* of the data to the next layer, the *bridge layer*, follows. The complete processing logic of the *data ingestion layer* can be seen in Figure 4.5.

4.3.2.3 Bridge Layer

The proposed system reflects the general IoT architecture described by Mishra *et al.* where the *bridge layer* constitutes the entry point into the core system [22] and is running on the central processing unit because the data does not need to leave the local network in this case. It logically links the *data ingestion layer* (producer) with the *processing layer* (consumer) and serves as the intermediary communication hub, see Figure 4.6. The decoupled architecture design allows for asynchronous and scalable data transfer [148].

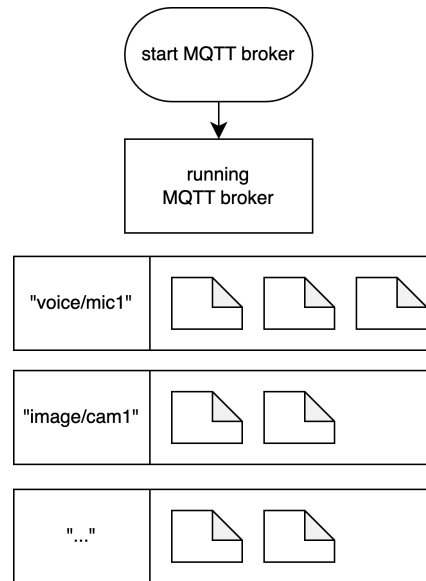


Figure 4.6: Process logic of *bridge layer* – flow diagram.

The *bridge layer* is implemented as a (formerly) Message Queueing Telemetry Transport Protocol (MQTT) broker by *Mosquitto* running in a Docker container [187]. Quality characteristics such as simplicity, openness, lightweightness, and easy deployment make MQTT the ideal fit for constrained environments such as the one in this proposed system [22]. MQTT consists of three participants: producer, broker, and consumer. While both – producers and consumers – are MQTT clients, the broker is an MQTT server [22]. MQTT offers three "quality of service" (QoS) specifications for the delivery of

messages: "(0) At most once", "(1) at least once", and "(2) exactly once" [22]. In this system it is negligible if single messages are not correctly received or if duplicates of messages occur. Therefore, the system implements the QoS "(0) at most once" because this QoS is the easiest, quickest, and cheapest option among them [22]. Note that instead of defining the QoS at the broker level, the individual producers/consumers define it when publishing/consuming on certain topics, as this leaves open the possibility of better QoS guarantees for critical data.

MQTT allows fine-grained routing of data streams via topics. The *data ingestion layer* defines, during the *transport* phase, the topic to which the collected and filtered data should be routed. It is recommended to use its own dedicated topic for each sensor input type. This fine-grained separation enables the potential for reorganizing data streams to find meaningful and insightful combinations. Once data is published on a given topic, multiple consumers may consume it and use it according to their requirements. Figure 4.6 illustrates, how the *bridge layer* routes between producers and consumers conceptually. If a new sensor type is deployed in the system, the user can simply define the topic in the *data ingestion layer*, and the topic will automatically be created in the MQTT broker.

4.3.2.4 Processing Layer

The *processing layer* consists of two different types of services: (1) *metrics computation* services and (2) *user profiling* services. As defined in the service-level architecture overview, see Table 4.3, both service types are intended to offer great modification and extensibility capabilities. Therefore, they both follow a ports and adapters service-level architecture paradigm. Figure 4.7 illustrates the process logic of this layer.

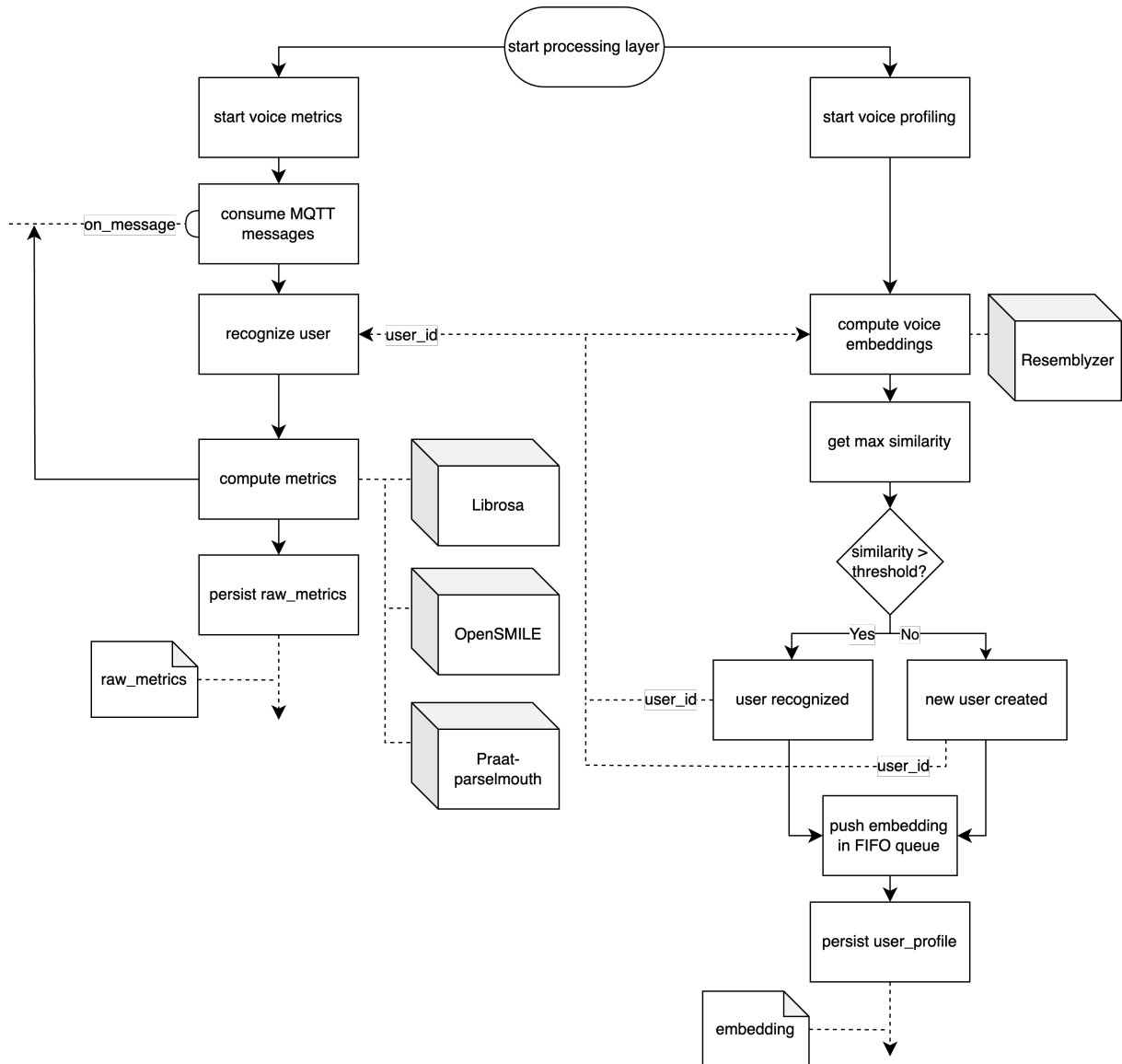


Figure 4.7: Process logic of *processing layer* – flow diagram.

Metrics computation: These services are the consumers of the MQTT topics, process the raw data in order to compute metrics based on it, and store the metrics in the *persistence layer*. The *metrics computation* services depend on the *user profiling* services in order to obtain the *user_id* to which a given data segment belongs. This introduces tight coupling; however, both types of services are within the same semantic group of the *processing layer*.

The functional requirement *FR-10.3*, see Section 4.2.2, defined that *metrics computation* services must be capable of consuming data streams from multiple MQTT topics at the same time to aggregate them for meaningful metrics. To support this, the system proposes a ports and adapters strategy, where individual *Handlers* can be registered with

adapters and are tailored specifically to handle different topics for a given *UseCase*. The *ConsumerPort* specifies the default behavior of any kind of data input logic, *i.e.*, adapters must provide a *register_handler(handler)* method where handlers can be added. Because the implementation uses MQTT as a messaging broker in the *bridge layer*, this port is implemented as a *MqttConsumerAdapter*. This adapter is implemented in such a way that the MQTT client is subscribed to the set of all topics of his registered handlers. When a message arrives, the adapter forwards the message to the respective handlers that are registered to this particular topic [128].

The logic for the *metrics computation* is placed at the core of this service. This computation is implemented according to the specifications in Section 3.4 and Section 3.5.1. To accomplish this, several external tools were used. Within the *metrics computation* logic, the individual tooling methods are invoked, including functions from *OpenSMILE* [62], *Praat(-parselmouth)* [177], and *Librosa* [21]. Each metric is initially represented as a flat dictionary and later individually inserted into the database.

User profiling: These services are responsible for recognizing the user to whom the given data segment belongs. As the architecture follows a multimodal approach, there is a need for multiple *user profiling* services, each responsible for one modality. This is reflected through the definition of ports and adapters of different modalities. For example, when a *metrics computation* service processes raw audio data, it can send a request to a *user profiling* service that implements the *UserRecognitionAudioPort* – such as the concrete *voice_profiling* service – which then retrieves the *user_id* for the given audio data [128]. As the ports and adapters describe the contract and define what can be expected by the service, the dependency never breaks, and services can be exchanged easily to try out different *user profiling* techniques or services.

The implemented *user profiling* service, *voice_profiling*, receives raw audio data, which is then processed with the *VoiceEncoder* of *Resemblyzer* [198]. *Resemblyzer* is a Python library for analyzing and comparing voices. It extracts high-level embeddings in a vector representation from tiny voice samples using a deep learning model [198]. First, *Resemblyzer* is used for speaker verification, using voice embeddings to identify the *user_id* of the given audio data. As there is a possibility of multiple people talking at the same time, another feature of *Resemblyzer* is used. With speaker diarization, the audio signal is segmented by the current speaker. If, in the audio data, multiple segments are present due to overlapping user speech, these audio data are rejected, as the analysis of these audio data would falsify the values produced by the *metrics computation*.

The implemented *voice_profiling* service works as follows: At startup all voice embeddings are loaded from the *persistence layer*. Upon request, a new embedding of the received audio signal is created, which is then compared with the existing user voice embeddings. If a match above the defined threshold of 0.7 has been found, the user has been identified, the user voice embedding updated with the newly computed voice embedding, and the *user_id* is returned. In the case that no user voice embedding reaches the threshold, a new user profile is created, the user voice embedding is stored, and the new *user_id* is returned. If the speaker diarization detects multiple speaker segments, the service rejects the audio data and returns the respective status code to the *metrics computation* service [128].

4.3.2.5 Persistence Layer

As defined in the high-level architecture concept, see Section 4.3.2.1, the data ownership and data access are designed towards one single database instance containing multiple collections, where, respectively, one service is responsible for the writing operations, but multiple services can access the data simultaneously. Each collection can be associated with the individual phases of the data pipeline, making the preparation as well as the persistence of the data cumulative and easily interpretable. See Figure 4.8 for the detailed database model, data ownership, and data access relationships. In this implementation, the following collections are described [128]:

- *raw_metrics*: Contains all metrics of any modality that are directly derived from the raw data.
- *voice_profiling*: Stores the embeddings of users, obtained by the *voice_profiling* service.
- *aggregated_metrics*: Based on the records in the *raw_metrics* collection, cumulatively stores the daily aggregated metric values.
- *contextual_metrics*: Based on the records in the *aggregated_metrics* collection, cumulatively stores the contextual metric values.
- *analyzed_metrics*: Based on the records in the *contextual_metrics* collection, cumulatively stores the analyzed metric values.
- *indicator_scores*: Based on the records in the *analyzed_metrics* collection, cumulatively store the DSM-5 indicator scores.

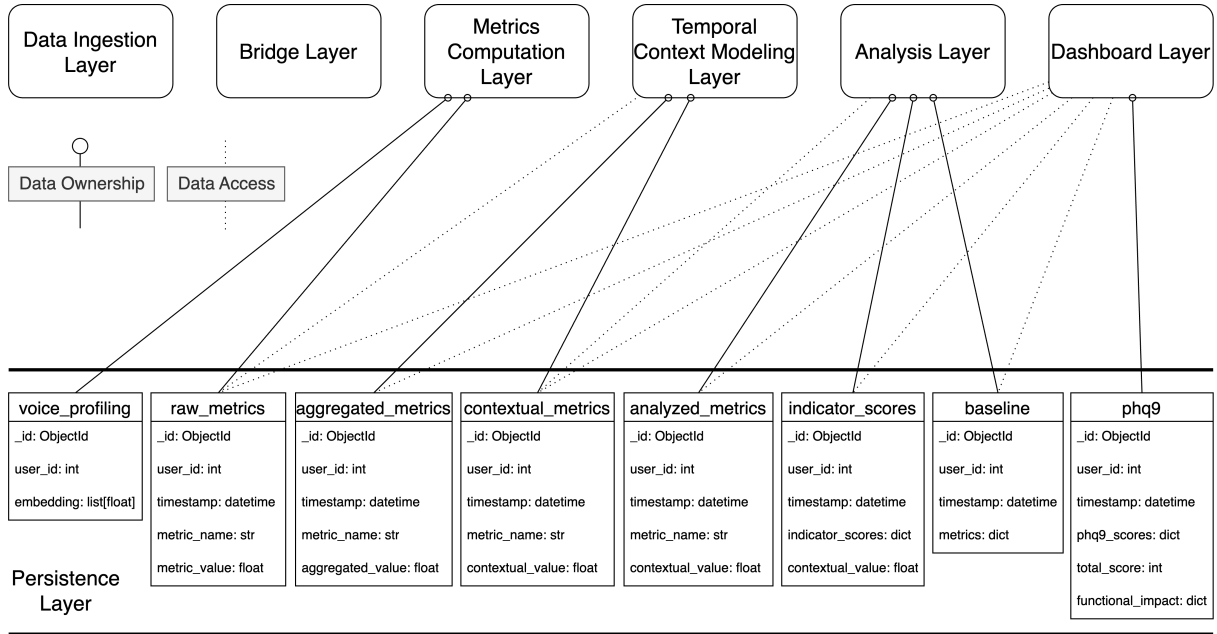


Figure 4.8: Implemented database collections model.

- *baseline*: Contains, for every metric, the baseline values (average and standard deviation) for the users in the system.
- *phq9*: Stores the submitted answers of the PHQ-9 questionnaire for the users in the system.

The database model shows that there are two dominant approaches employed for storing the data. In the collections that store metric data that has been captured and processed by the system (*raw_metrics*, *aggregated_metrics*, *contextual_metrics*, and *analyzed_metrics*), are stored as one record per metric. This enables flexible querying and efficient writing operations due to the granular nature of the records. This also enables the easy integration of other modalities through different deployed sensors in the environment. On the other hand, the collections *indicator_scores*, *phq9*, and *baseline* store the aggregated values as a single record, due to the semantic unity of the individual values. This makes writing and query operations easier to handle.

4.3.2.6 Temporal Context Modeling Layer

The next phase of the data pipeline is the *temporal context modeling layer*, which transforms raw metric data into a temporally contextualized representation that reflects the user's dominant behavioral and physiological patterns over time. This layer is implemented as illustrated in Figure 4.9 and is designed to be rigid – robust against

outliers – ensuring that anomalous data points do not distort the overall metric trends. Instead of treating each metric observation in isolation, the model incorporates historical values, assigning greater influence to more recent data while gradually diminishing the weight of older entries. This temporal weighting allows the system to capture meaningful trends and behavioral continuity, providing a stable and context-aware interpretation of the user's state.

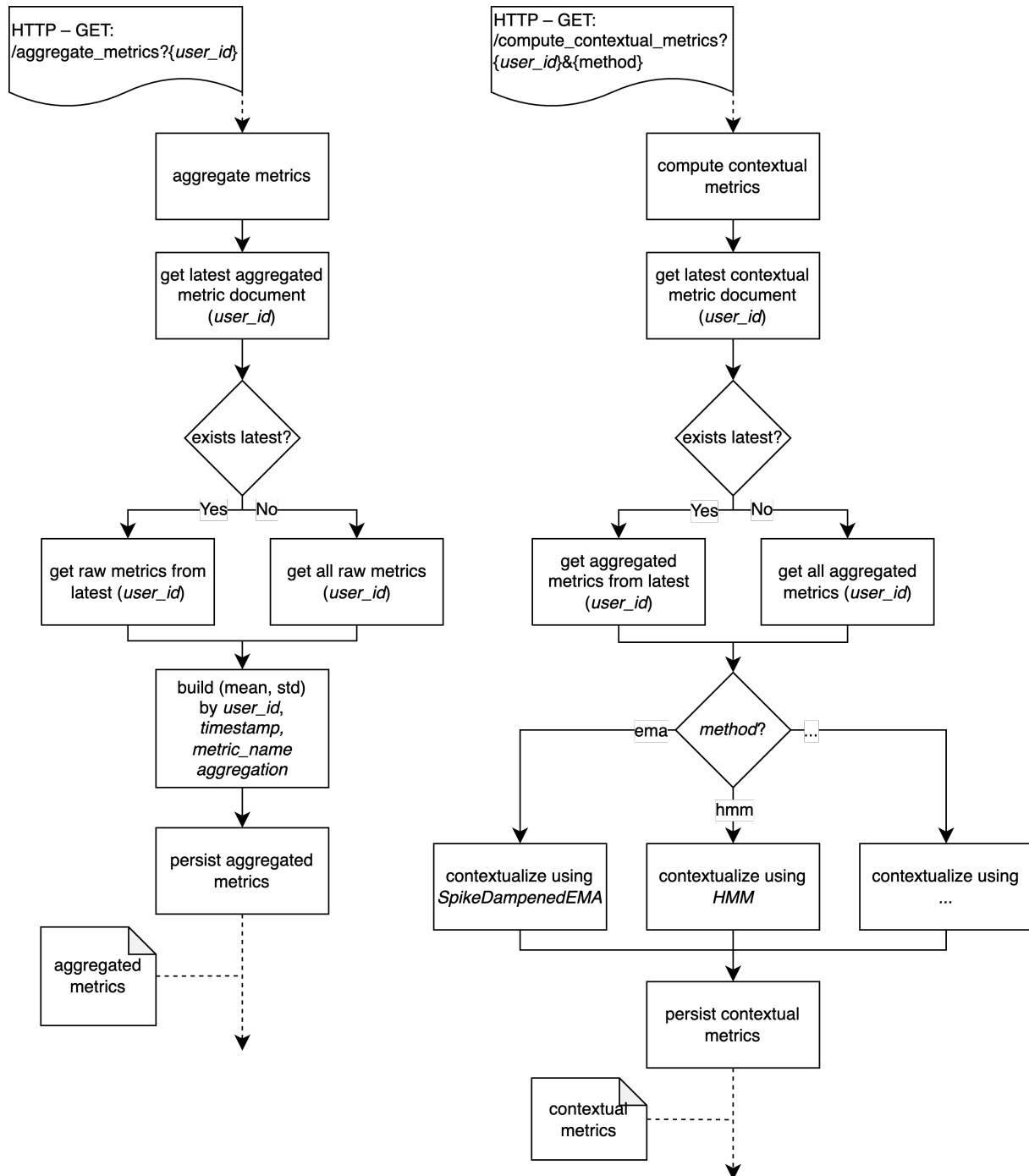


Figure 4.9: Process logic of *temporal context modeling layer* – flow diagram.

To achieve this behavior in a resource-efficient manner, the system employs a two-fold process. First, *raw_metrics* are summarized into daily aggregates, based on the assumption that automated depression detection does not require a fine-grained intraday analysis. Instead, it is assumed that daily aggregates are sufficient to capture meaningful behavioral and physiological trends. In addition, in this layer, the ports and adapters architectural design allows the user to easily change or extend the aggregation process. By default, the system computes the average value of each metric per day, but this behavior can be customized by implementing an alternative *AggregateMetricsUseCase* [128]. Once aggregated, the daily metrics are persisted in the *aggregated_metrics* MongoDB collection [158].

Second, the *aggregated_metrics* are contextualized on the temporal axis. For contextualization of aggregated data, there exist many different options to pursue. For extensibility, an abstract class *Contextualizer* is provided, where users can implement their own algorithms for the *temporal context modeling* [128].

Hidden Markov Model (HMM): One possibility is to use an HMM, which is a statistical model that can describe the evolution of data points. HMMs are particularly popular in sequence analysis problems, such as speech recognition or DNA sequencing problems [247]. Especially context-sensitive HMMs (csHMM) are of interest due to their capabilities in using parts of data points in the past (context) to adjust the probabilities at future states. This is useful in scenarios where the goal is to describe long-range correlations between states [247]. The implementation of the system offers a basic HMM implementation for capturing the temporal context of the *aggregated_metrics*.

Exponential Moving Average (EMA): Another way to model the temporal context of *aggregated_metrics* is an EMA. EMA is a much simpler method compared to HMM, but suffices to achieve the desired goal in this scenario. The EMA provides a way to track trends or detect changes in time series data. This algorithm gives more weight to recent data points, as recent events tend to be more predictive of the near future [185]. The EMA algorithm is computationally efficient, as this recursive formula, see Algorithm 4.1 and Algorithm 4.2, only needs the current value and the last smoothed value of the time-series data to compute the next *contextual_metric* [28].

The proposed implementation in this system uses a spike-dampened EMA variation, which filters out random noise by further dampening outliers (spikes) while still reacting to sustained changes rather quickly [28]. Those characteristics of the spike-dampened EMA are particularly useful in the household environment where sensor data is likely to

Algorithm 4.1 Smoothing factor for EMA [28].

INPUT: N : Number of periods for smoothing**OUTPUT:** α : Smoothing factor for EMA, with $0 < \alpha \leq 1$ **Compute:** $\alpha = \frac{2}{N+1}$

Algorithm 4.2 EMA update computation [28].

INPUT: x_t : Observation at time t EMA_{t-1} : EMA value at time $t - 1$ α : Smoothing factor, $0 < \alpha \leq 1$ **OUTPUT:** EMA_t : Updated EMA at time t **Compute:** $\text{EMA}_t = \alpha \cdot x_t + (1 - \alpha) \cdot \text{EMA}_{t-1}$

be noisy. Algorithm 4.1 and Algorithm 4.2 are the same for the spike-dampened EMA [28] with the addition of spike damping; see Algorithm 4.3, where x'_t is the dampened input. Note that spike-dampened EMA is not a standardized algorithm, but the concept of damping spikes is useful in this scenario of highly volatile data.

Algorithm 4.3 Variation – spike-dampened EMA computation.

INPUT:

x_t : Observation at time t

EMA_{t-1} : EMA at time $t - 1$

δ : Spike threshold (maximum allowed deviation from EMA before dampening)

γ : Fixed dampening factor

OUTPUT:

x'_t : Dampened observation at time t

if $|x_t - \text{EMA}_{t-1}| \leq \delta$ **then**

$x'_t \leftarrow x_t$

else

$x'_t \leftarrow \text{EMA}_{t-1} + \text{sign}(x_t - \text{EMA}_{t-1}) \cdot \gamma$

return x'_t

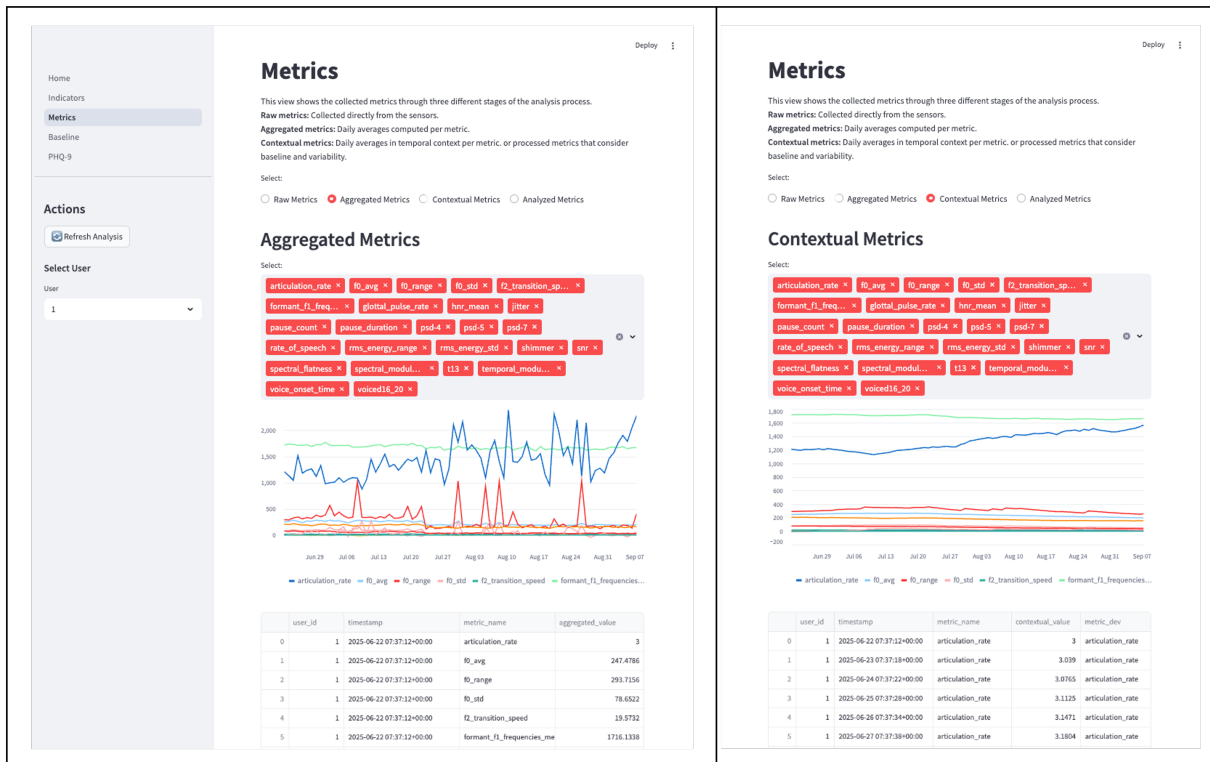
Applying these concepts, the following parameters have been defined [128]:

- *Smoothing factor*: According to the DSM-5 specification, five (or more) clinical indicators must be present during the same two-week period [18]. To reflect this temporal requirement in the *contextual_metrics*, a smoothing factor of 0.13 has been chosen. Based on the formula for the EMA smoothing factor, see Algorithm 4.1, this value approximates the effect of a two-week observation window, simulating the intended temporal span of the DSM-5 specification.
- *Spike threshold*: As the collected data and corresponding metrics vary significantly in magnitude, applying a single, hard-coded spike threshold is not smart. Instead, a dynamic approach is employed, where the spike threshold is set to 10% of the currently observed data point. However, a better approach could involve determining the spike threshold relative to the user's current baseline.

- *Dampening factor*: The dampening factor of 0.3 has been shown to produce the best results.

The processing of the metric values is performed in a continuously cumulative manner. This approach is based on the assumption that data is collected and persisted on the same day. By processing the data cumulatively, the system ensures that all relevant raw metric data – including newly added data from additional sensors – is incorporated into the corresponding metrics. This cumulative strategy is not limited to any phase of the data pipeline and is consistently applied throughout all stages of the data pipeline. In addition to ensuring data completeness and consistency, it also offers resource efficiency, as only new data needs to be processed rather than computationally expensive recomputations from scratch.

Using the *depressed* (*long_depressed_sample_nobreak.wav*) and the *non-depressed* (*long_nondepressed_sample_nobreak.wav*) sample data, Figure 4.10 shows both steps, *aggregate metrics* and *contextualize metrics*, involved in the *temporal context modeling* layer. See GitHub for the sample data files [128]. See Appendix B for high-resolution screenshots.



4.3.2.7 Analysis Layer

The last and final step in the data pipeline consists of the transformation of the *contextual metrics* into the nine individual DSM-5 *indicator scores*. As described in Chapter 3, the selected features are grouped by their associative relationship to depression indicators and their respective analysis methodologies. The transformation process is divided into two parts: (1) the *contextual metrics* are analyzed by quantifying the deviation of the observed value from the user's baseline, and (2) the results are evaluated based on the defined associative mappings to build the *indicator scores*. Currently, the system supports only linear relationships; analysis of non-linear relationships requires additional methods to be developed and integrated.

(1) Metrics Analysis

The process logic of the sub-process *metrics analysis* is illustrated in Figure 4.11.

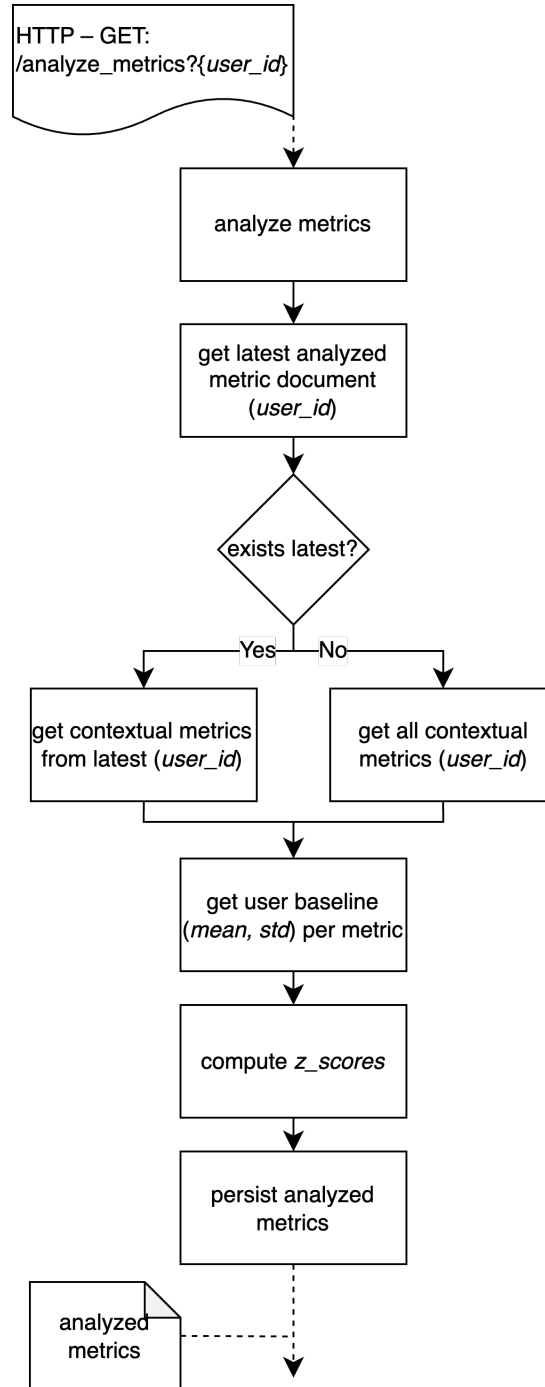


Figure 4.11: Process logic of *analysis layer*, sub-process *metrics analysis* – flow diagram.

The analysis of the *contextual metrics* relies on quantifying how much an observed value deviates from a user’s typical behavior. The work of Saha applied a standard scaling approach using z-scores, as it makes sense to normalize the features to ensure equal contribution during the computation of the individual DSM-5 indicator scores [49]. This is important because the proposed algorithm is sensitive to feature scales. In this master’s thesis, the same approach is followed: The z-scores are computed according to Algorithm

4.4 [220], which measures the number of standard deviations an observation differs from the user’s baseline mean [220], for example, see Figure 4.12.

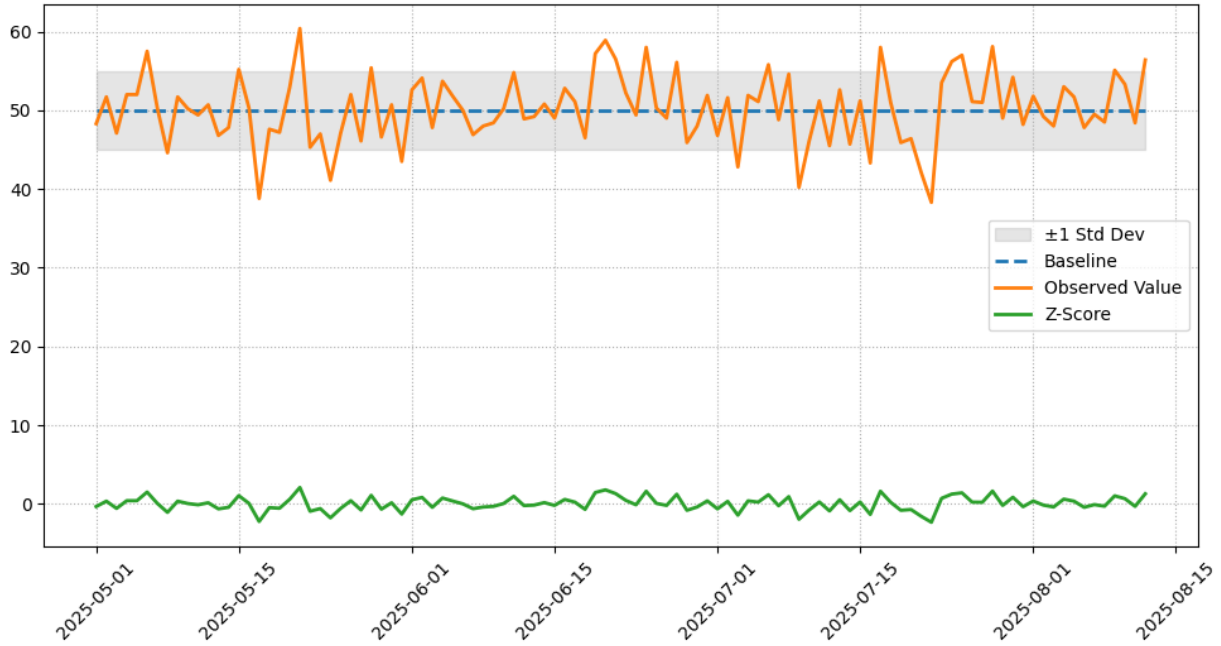


Figure 4.12: Z-score computation – explanatory visualization [220].

To compute these scores, two inputs are required: the current *contextual metric* values (output from the *temporal context modeling layer*) and the user’s baseline statistics. Since no ground truth is available to the system at the initialization, multiple strategies for approximating the user’s baseline must be evaluated.

Algorithm 4.4 Z-score computation [220].

INPUT:

- x : Contextual metric value
- μ : Baseline mean
- σ : Baseline standard deviation

OUTPUT:

- z : Z-score (standardized value)

$$z \leftarrow \frac{x - \mu}{\sigma}$$

return z

One approach could be to assume that the user's behavior of the initially collected data reflects their typical state. This method enables an easy integration of new users with quick adaptation to the user's individual behavior but risks incorporating already anomalous behavior into the baseline, which could result in wrong DSM-5 indicator scores. For example, if the user is already depressed, the system would detect depression if the mental state of the user gets worse. An alternative approach – which the implementation follows – includes the possibility to upload baseline files *e.g.*, *population_baseline.json*. Baseline records are persisted to the database. The step *get user baseline (mean, std) per metric* from the flow diagram in Figure 4.11 retrieves the last persisted baseline record for the given *user_id*. When a new user in the system is initialized, no baseline record can be found in the database. In this case, the baseline values from the uploaded baseline file are used. If a user already has a baseline record persisted in the database, all those values are used and are given priority over the values from the baseline file [128].

(2) DSM-5 Indicator Derivation

The process logic of the sub-process *DSM-5 indicator derivation* is illustrated in Figure 4.13.

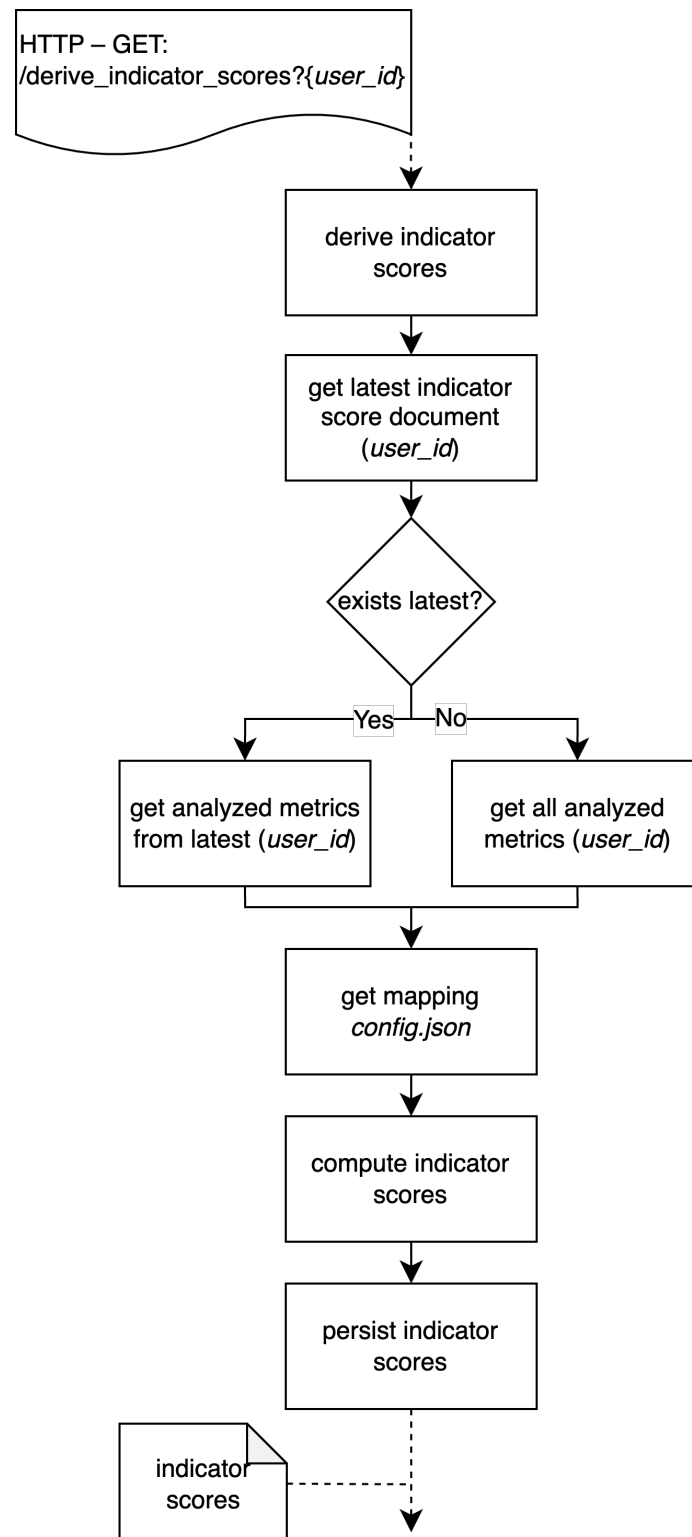


Figure 4.13: Process logic of *analysis layer*, sub-process *indicator derivation* – flow diagram.

Once the z-scores have been computed, the DSM-5 *indicator derivation* takes place. The conceptual feature mapping is directly applied in this part of the system. At the

initialization of the system, the concrete *Mapping Pyramid* is read from the mapping configuration file, *i.e.*, *config.json* [128]. As scientists may investigate different feature mappings to validate their hypotheses of effective feature mappings for depressive behavior (or other mental health disorders), they can store their own mapping configuration file inside the *mapping* directory. This is a flexible, easy, and efficient way to completely redefine the interpretation of metrics in terms of the individual DSM-5 indicators. The structure of the mapping configuration file is straightforward; see Figure 4.14: Each DSM-5 indicator is represented as a top-level key (*e.g.*, *1_depressed_mood*) and contains a *metrics* object. This object lists all metrics that are mapped to this particular indicator. Each metric provides the corresponding metadata for the fine-grained mapping logic, as specified by the *Feature Mapping Specification*, see Section 3.5.

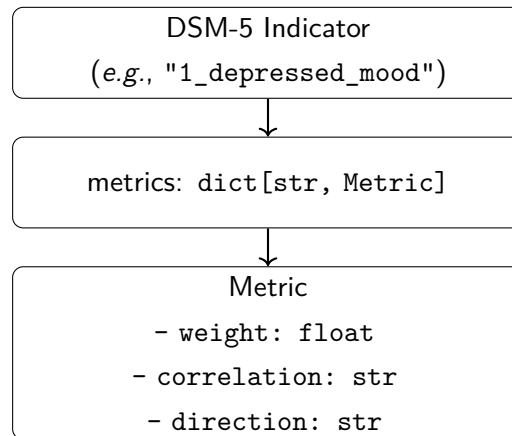


Figure 4.14: Syntactic file structure of the *config.json* file [128].

In the computation of the DSM-5 indicator scores at time t , the *analyzed_metrics* at time t , the mapping configuration, as well as the latest (if present) DSM-5 *indicator_scores* are needed. The detailed computation is illustrated in Algorithm 4.6. For the final score for indicator i at time t , the weighted sum over the z-scores for every mapped metric is computed. Before building the individual indicator sums, the direction modifier is applied to the z-score, and a lower bound of 0 is imposed. If for any particular indicator no mapped metric values at time t are present, the latest available *indicator_score* is used. If no latest *indicator_score* is available, the indicator is evaluated with the value *None* [128].

Ultimately, the DSM-5 *indicator_scores* are also persisted in the database, as those scores constitute the final step in the data pipeline of the system.

Algorithm 4.5 Direction modifier function.

INPUT: z : Z-scoredirection $\in \{\text{positive, negative, both}\}$ **OUTPUT:** $f_d(z)$: Modified z-score based on direction**if** direction = positive **then** $f_d(z) \leftarrow z$ **else if** direction = negative **then** $f_d(z) \leftarrow -z$ **else if** direction = both **then** $f_d(z) \leftarrow |z|$ **return** $f_d(z)$

Algorithm 4.6 Indicator score computation.

INPUT: M_i : Set of metrics associated with indicator i w_m : Weight for each metric $m \in M_i$ $z_m(t)$: Z-score for metric m at time t f_d : Direction modifier function**OUTPUT:** $S_i(t)$: Indicator score at time t $numerator \leftarrow 0$ $denominator \leftarrow \sum_{m \in M_i} w_m$ **for each** $m \in M_i$ **do** $score \leftarrow \max(0, f_d(z_m(t)))$ $numerator \leftarrow numerator + w_m \cdot score$ $S_i(t) \leftarrow \frac{numerator}{denominator}$ **return** $S_i(t)$

Baseline Finetuning

The process logic of the sub-process *baseline finetuning* is illustrated in Figure 4.15.

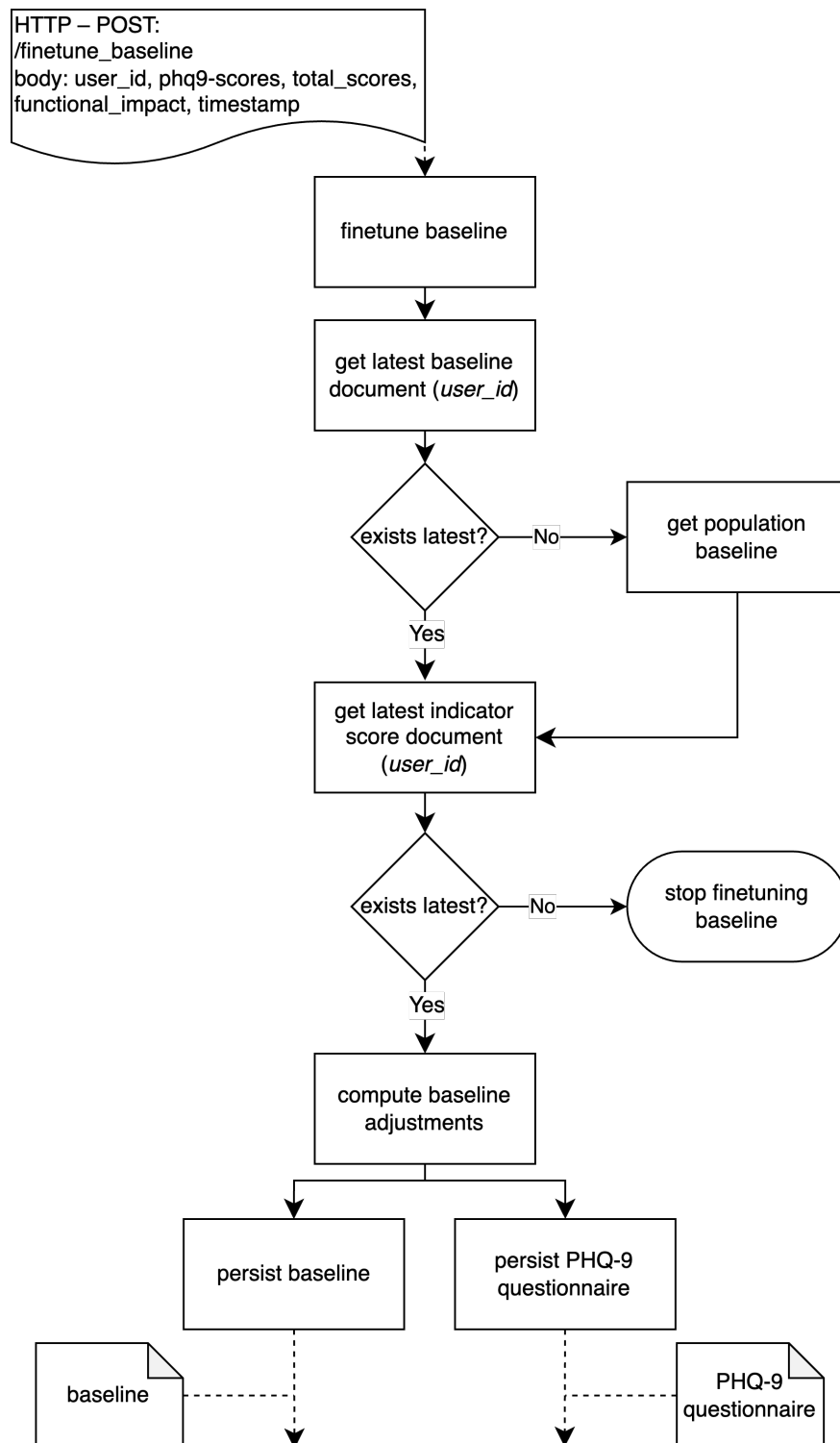


Figure 4.15: Process logic of *analysis layer*, sub-process *baseline finetuning* – flow diagram.

When the *contextual_metrics* are analyzed, they are evaluated against the user's baseline. In the proposed implementation, the initial baseline is defined by a user-specified *population_baseline.json* file [128]. However, as user behavior may evolve, particularly in cases such as children undergoing puberty or individuals experiencing significant life changes, there must be a mechanism to adapt the user's baseline over time. The proposed implementation for this mechanism follows a form of weakly supervised adaptation. To fine-tune the baseline, the user can complete the PHQ-9 questionnaire [109] and submit the data to the *analysis layer* via the form that can be found in the user-facing *dashboard*. The direct connection between the PHQ-9 answers and the baseline of the user on the metric level is evident. While the PHQ-9 questionnaire evaluates the subjective severity degrees on the DSM-5 indicator level [109], the outputs of the system represent the semantic counterpart: A direct comparison between system output and the provided user feedback can be performed.

The detailed computation of the updated baseline mean is shown in Algorithm 4.7. The baseline adjustment is based on the difference between the actual PHQ-9 scores and the latest DSM-5 *indicator_scores*. The adjustments are quantified by the metric's metadata *direction* and *weight*, which can be obtained from the mapping configuration (*config.json*). The learning rate η controls the step size of the baseline adjustments, making the baseline selectively more or less reactive to user feedback. Finally, baseline adjustments align the perceived severity of the disorder with predicted *indicators_scores*, thus improving the precision of the overall system [128].

Using the *depressed* (*long_depressed_sample_nobreak.wav*) and the *non-depressed* (*long_nondepressed_sample_nobreak.wav*) sample data, Figure 4.16 shows both steps, *analyze metrics* and *indicator derivation*, involved in the *analysis layer*. See GitHub for the sample data files [128]. See Appendix B for high-resolution screenshots.

Algorithm 4.7 Baseline finetuning function.

INPUT:

$\mu_m^{(t)}$: Current baseline mean of metric m
 σ_m : Baseline standard deviation of metric m
 y_i : Actual PHQ-9 score for indicator i
 $\hat{y}_i$: Predicted DSM-5 score for indicator i
 $w_{i,m}$: Weight of metric m for indicator i
 η : Learning rate (e.g., 0.2)
 $f_m(\cdot)$: Direction modifier for metric m

OUTPUT:

$\mu_m^{(t+1)}$: Updated baseline mean
 $error \leftarrow y_i - \hat{y}_i$
 $adjustment \leftarrow \eta \cdot f_m(error \cdot \sigma_m) \cdot w_{i,m}$
 $\mu_m^{(t+1)} \leftarrow \mu_m^{(t)} + adjustment$
return $\mu_m^{(t+1)}$

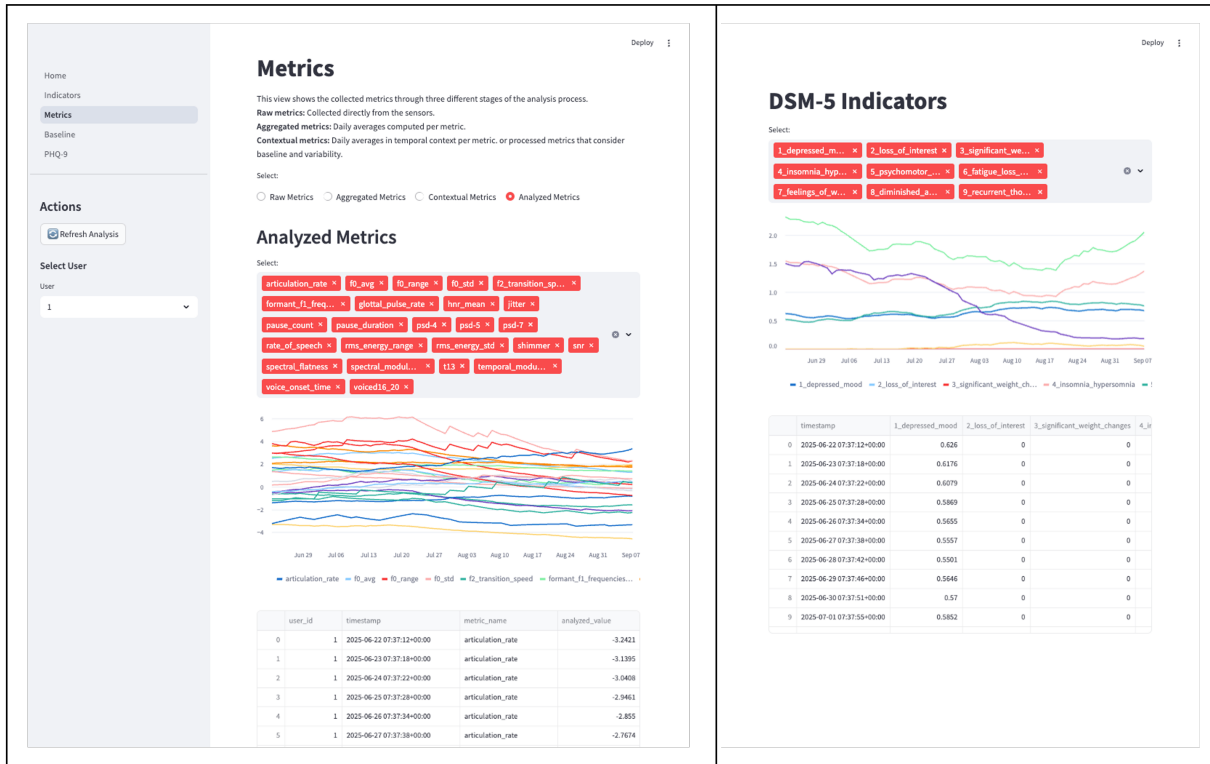


Figure 4.16: Process steps – (left) *analyze metrics* and (right) *indicator derivation*.

4.3.2.8 Dashboard Layer

Finally, the *dashboard layer* serves as the interface through which users interact with the system. The implemented *Stream-lit dashboard* application [88] provides the user with detailed insights into the data along the data pipeline, making it easy to gain insights about the data as well as facilitating reasoning about the system's outputs.

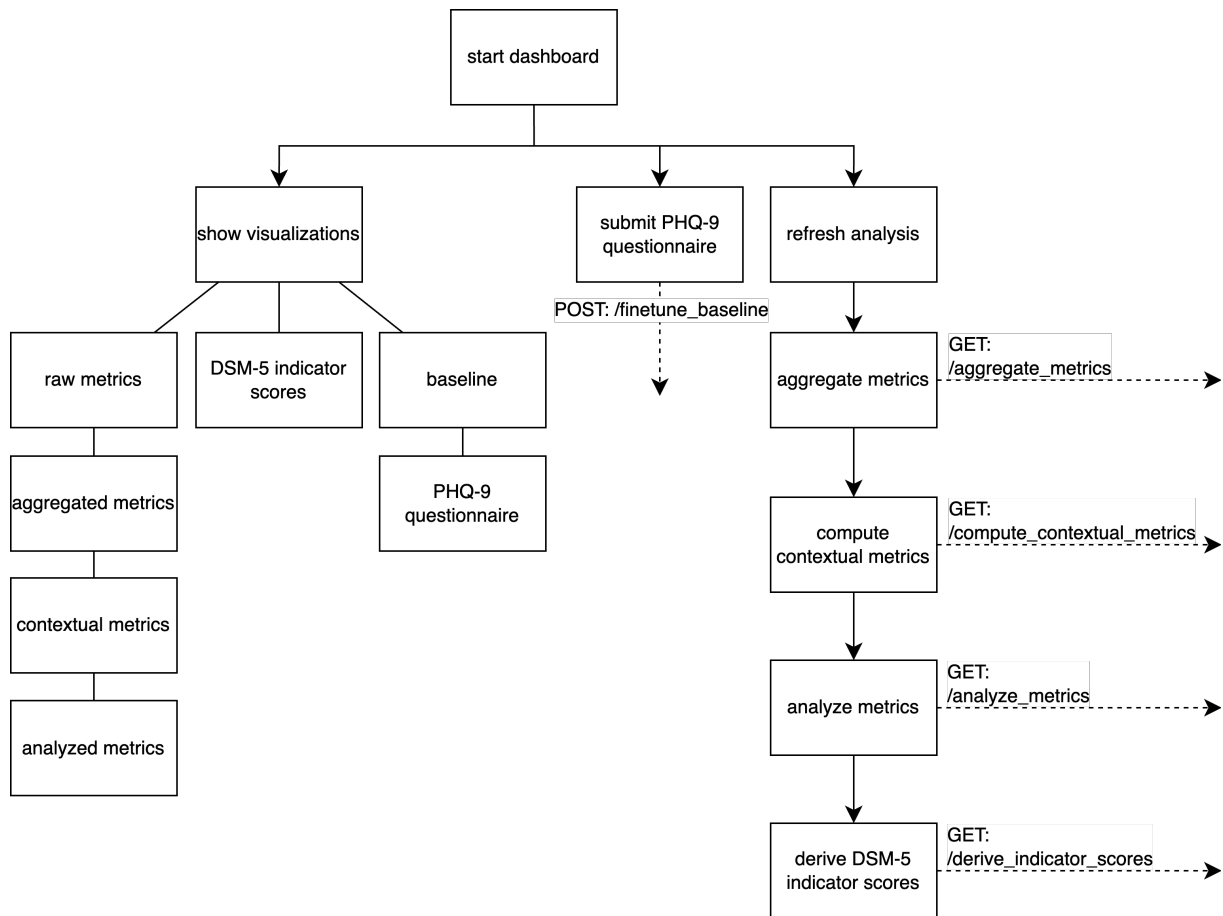


Figure 4.17: Process logic of *dashboard layer* – flow diagram.

The *dashboard* orchestrates the individual components, see Figure 4.17, of the data pipeline and presents them in a simple way to the user. Additionally, the *dashboard* enables users to complete the PHQ-9 questionnaire and submit their responses to the *analysis layer* in order to finetune the user baseline. Illustrations of the *dashboard* interface are shown in Figure 4.18 and Figure 4.19; see Appendix B for high-resolution screenshots.

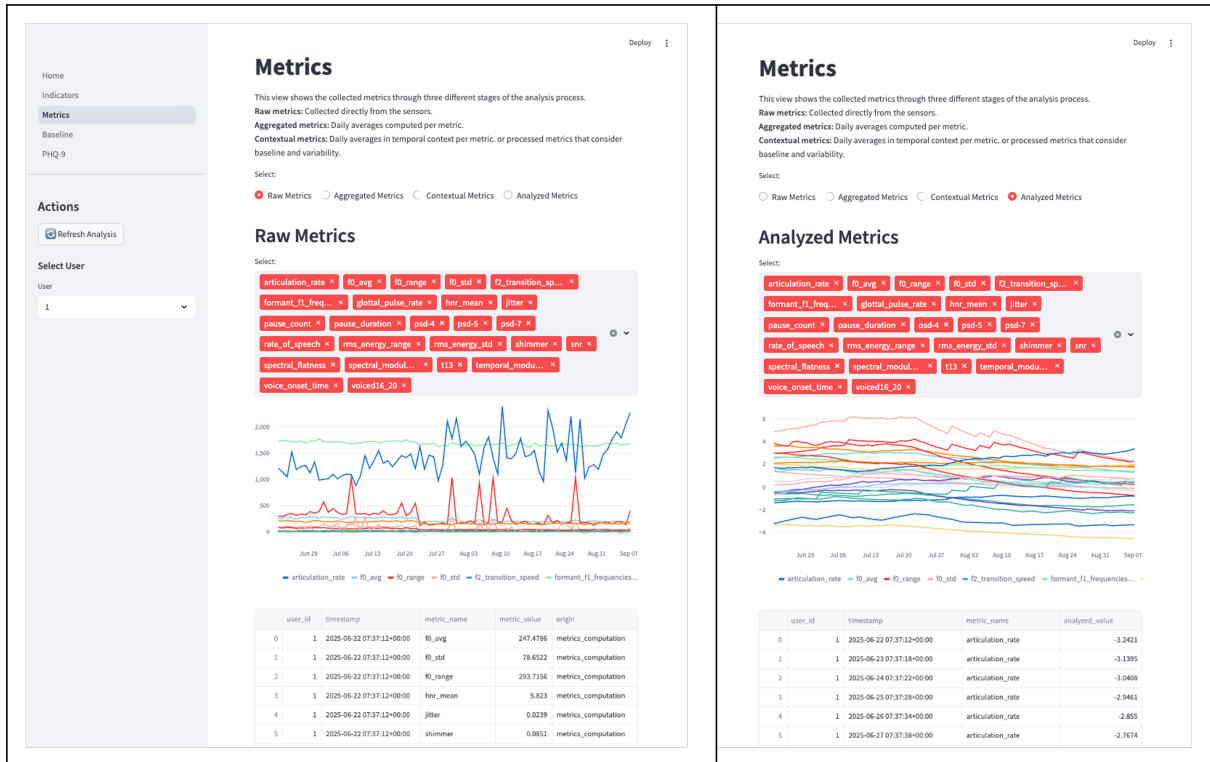


Figure 4.18: Dashboard screenshots – (left) raw metrics and (right) analyzed metrics.

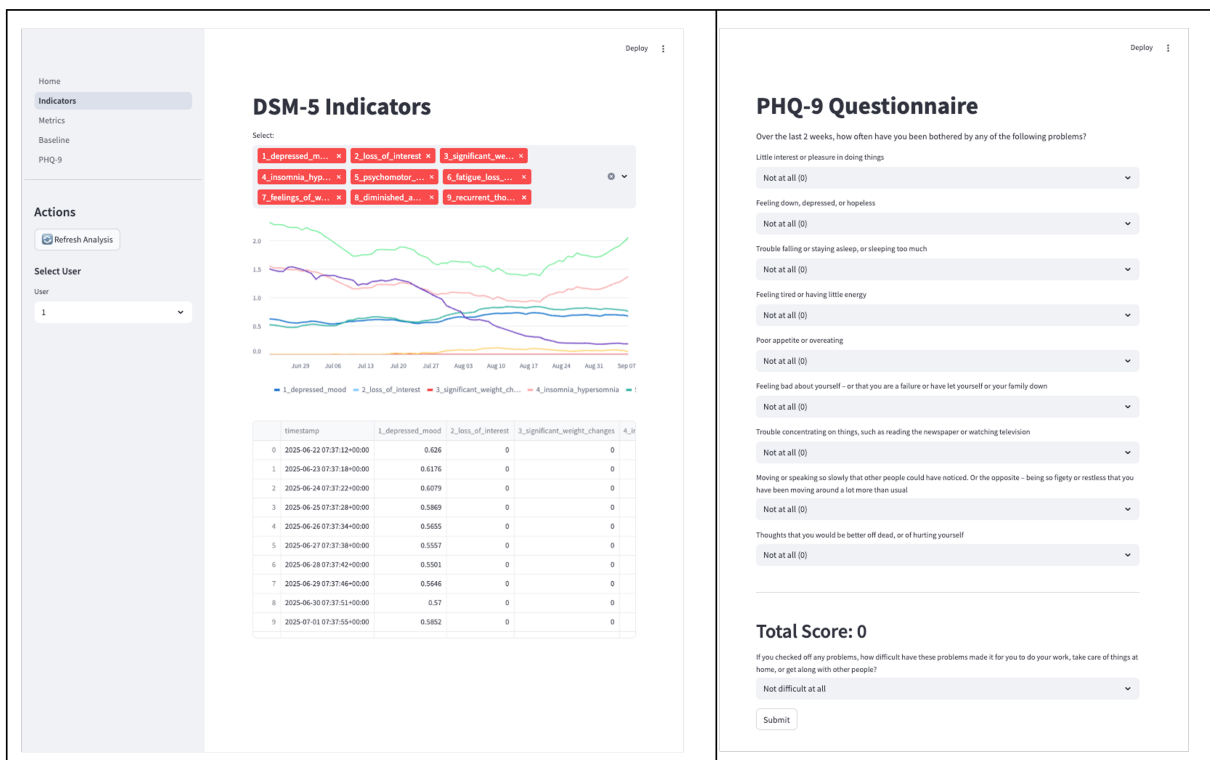


Figure 4.19: Dashboard screenshots – (left) indicator scores and (right) PHQ-9 questionnaire.

Chapter 5

Evaluation

In the evaluation of this master’s thesis, each objective described in Table 1.1 is inspected in detail to determine to what extent the overarching research question can be answered.

How can an explainable white-box system for automated health assessments in IoT-enabled private household environments be designed while promoting scientific collaboration and innovation; and to what extent can acoustic-based voice features – alone or in combination with other modalities – be integrated into that system to promote automated depression detection?

Obj-4 is addressed in this chapter about the evaluation, where the validity of each contribution of this work is examined.

5.1 System Architecture

The *System Architecture* (*Obj-1*) has been proposed in Chapter 4, together with the prototypical instantiation in Section 4.3. A two-fold approach is followed: First, the *System Architecture* is evaluated based on its ability to meet the functional requirements defined in Section 4.2.2. Second, the *System Architecture* is assessed based on its ability to fulfill the non-functional requirements as defined in Section 4.2.3. Note that user stories as defined in Section 4.2.1 are not evaluated, as their primary purpose was to guide the derivation of detailed functional and non-functional requirements. As such, evaluating them would not reveal additional insights. This hybrid approach combines binary and qualitative assessment methods to provide a more holistic evaluation of the proposed *System Architecture*.

5.1.1 Functional Assessment

Based on the prototypical implementation that is an instantiation of the *System Architecture*, the evaluation of functional requirements focuses only on the coverage of whether each requirement has been implemented in any form. An evaluation of the quality of

the overall implementation follows in later sections of this chapter, see Section 5.1.2. The presented checklist method is used, where each requirement is marked as: *Fully implemented* – [✓], *Partly implemented* – [(✓)], *Not implemented* – [–], and *Rejected* – [×]. Additional notes clarify details and reference relevant parts where necessary. The following Table 5.1 shows the evaluation of the *System Architecture* according to these functional requirements.

Table 5.1: Checklist of implemented functional requirements.

Nr.	Description	Done	Note
FR-1	The system must provide an installation script for the system core that sets up all dependencies and services automatically. Setup requires (1) running an installation script on the server and (2) starting a script on each sensor device.	✓	The microservices are configured and deployable via Docker compose, see <i>docker_compose.yml</i> file.
FR-2	At startup, the system must load default configuration settings, initialize background services, and begin collecting data.	✓	Start collecting data on the edge devices by the <i>sense.py</i> file.
FR-3	It must be possible for the user to modify the configuration file located in the installation directory.	(✓)	Possible but requires modifying Python code.
FR-4	The system must provide both a REST API and a web-based dashboard for accessing core functionalities.	✓	Both options are implemented. The orchestrated processes in the <i>dashboard</i> remove many complexities, making it much easier to use.
FR-5	The system must be accessible from any device within the same local network, with no authentication required, assuming a trusted private household environment.	✓	The system is freely accessible from within the local network.
FR-6	The system must provide access to raw data, contextualized data, analyzed data, and the DSM-5 indicator scores via the dashboard in an interpretable format.	✓	The data across all processing stages is available in both a tabular and a graphical representation.
FR-7.1	The system must continuously capture audio from configured microphone devices or simulated from an audio file.	✓	Available implementation for continuous audio capture and provided abstractions for additional input sensor types.
FR-7.2	Audio must be captured with a minimum sampling rate of 16 kHz.	✓	The sampling rate is set by default to 16 kHz. It can be adjusted at initialization of the respective edge device abstraction.
FR-8.1	Captured audio data must be streamed from sensor devices to the central processing unit.	✓	An intermediary MQTT broker is utilized to decouple data producers from the consuming entities.
FR-8.2	Audio data must be encoded in mono PCM and must include metadata such as the device and a timestamp.	(✓)	The field for the device is missing.

Nr.	Description	Done	Note
FR-8.3	Data must be available to multiple processing pipelines in parallel.	✓	Due to the intermediary MQTT broker, multiple <i>metrics computation</i> services may consume and process the same data.
FR-9.1	Preprocessing must be specifically tailored for each data modality and the respective downstream task.	✓	As of now, only speech data is analyzed by the system. The given abstractions in the <i>data ingestion layer</i> offer an easy integration of custom data preprocessing, see <i>AbstractEdgeDevice</i> and following.
FR-9.2	Audio preprocessing must at least include voice activity detection (VAD), segmentation into chunks of 1-5 seconds, voice recognition, and the rejection of noise.	✓	The preprocessing (voice activity detection, segmentation, voice user recognition, and rejecting noisy segments) is robust in a clinical setting.
FR-10.1	Metric computation must follow the specifications in the feature engineering section.	✓	The voice metrics are implemented according to the <i>Feature Mapping Specification</i> .
FR-10.2	Metrics must be associated with a specific user.	(✓)	Only voice user recognition is implemented. Other modalities require separate user recognition. In the future, relating a user between modalities may become challenging.
FR-10.3	It must be possible to aggregate multiple data streams of different modalities during the metric computation.	✓	The implemented <i>metrics computation</i> service, i.e., <i>voice_metrics</i> , implements the <i>MqttConsumerAdapter</i> interface and offers a way to consume and combine multiple data streams, i.e., <i>register_handler(handler)</i> .
FR-11.1	The system must maintain rolling user recognition profiles and trigger profile recalibration in case of persistent deviations.	✓	<i>User profiling</i> is implemented in such a way that the profile is continuously updated.
FR-11.2	A new user profile must be created if data cannot be matched to existing profiles.	✓	A new user profile is created if data does not exceed the similarity threshold of 0.7 for any of the user voice embeddings.
FR-12	Computed metrics must be stored in a central database with metadata such as the user and a timestamp.	(✓)	The field for the device is missing.
FR-13.1	Contextual metrics must be computed based on recent metric trends and aggregated daily.	(✓)	This is implemented in the <i>temporal_context_modeling</i> service, where implementations of the abstract <i>Contextualizer</i> model time windows in a customizable manner. The time window is not configurable at runtime and requires modifying Python code.
FR-13.2	Contextual metrics must be stored in a central database with metadata such as the user and a date.	(✓)	The field for the device is missing.
FR-14.1	Analyzed metrics must be computed based on daily contextual metrics.	✓	The analysis is based on a comparison of the user's baseline and the current <i>contextual metrics</i> .

Nr.	Description	Done	Note
FR-14.2	<i>Analyzed metrics</i> must be stored in a central database with metadata such as the user and a date	✓	The <i>analyzed metrics</i> are stored daily in the database with appropriate metadata.
FR-15.1	DSM-5 indicator scores must be computed based on daily analyzed metrics.	✓	The computation of the DSM-5 <i>indicator scores</i> is based on the <i>analyzed metrics</i> and reflects the semantic mapping that is provided via the <i>config.json</i> file.
FR-15.2	DSM-5 indicator scores must be stored in a central database with metadata such as the user and a date.	✓	The DSM-5 <i>indicator scores</i> are stored daily in the database with appropriate metadata.
FR-16	The system must support data ingestion from up to 30 sensor devices, streaming data in parallel.	(✓)	The cloud-centric approach plans a distributed computing approach, where the number of edge devices is not limited. Note that this was not tested.
FR-17	The system must support the timely processing of up to 30 concurrent data streams.	(✓)	The cloud-centered approach plans a distributed computing approach, where the number of services in the <i>processing layer</i> is not limited. Computing individual metrics can be very time-consuming. If a <i>metrics computation</i> service is responsible for calculating many complex metrics, it can become overwhelmed, resulting in processing times that exceed the rate at which data are received.
FR-18	The system must support semi-structured storage of textual records to accommodate diverse data modalities.	✓	The utilized MongoDB instance can store a document with different fields, data types, and structures without enforced schema, thus supporting semi-structured data storage.
FR-19	User data must be stored for up to 5 years to support meaningful historical data analysis.	✓	MongoDB persists data indefinitely by default. For guaranteed data integrity, data backups and replication need to be put in place.
FR-20	The system must provide visualizations for historical data at each stage of the data pipeline.	✓	Visualizations are available for each stage of the data pipeline, see Appendix B.
FR-21	The system must be packaged as a containerized platform for a potential future cloud deployment.	✓	The services are containerized using Docker, allowing future deployment on cloud platforms.
FR-22	The system architecture must promote the easy integration and exchange of edge devices, preprocessing, and analysis components.	✓	The modular design, together with the architectural style of ports and adapters, supports the modification and extension of system capabilities with minimal impact on existing components.
FR-23.1	Users must be able to mark data points as invalid to exclude them from analysis.	×	This feature is currently not implemented.
FR-23.2	The system must allow resetting the data of specific users.	×	This feature is currently not implemented.
FR-23.3	The system must allow a full data reset of every user.	×	This feature is currently not implemented.

Nr.	Description	Done	Note
FR-24	The web-based dashboard must include an interactive PHQ-9 form. Responses are used to fine-tune the user's mental health baseline, and they must be stored with metadata such as the user and a timestamp.	✓	The PHQ-9 questionnaire can be submitted on the <i>dashboard</i> , where it is subsequently persisted in the database. This triggers the fine-tuning of the user's baseline.

5.1.2 Non-functional Assessment

While the previous section provided a structured evaluation of the functional side of the *System Architecture*, this section focuses on the non-functional requirement assessment. Each non-functional requirement as defined in Section 4.2.3, is analyzed in isolation, each providing a unique view on individual quality aspects of the system. Since these requirements are qualitative in nature, the evaluation is based solely on reasoning.

Architectural flexibility (NFR-1): Architectural flexibility can be decomposed into scalability, modularity, and extensibility. Enabling horizontal scaling in the proposed system is specifically only addressed where it is needed. As horizontal scaling refers to the notion of replicating services in order to distribute the load among replicas, it is necessary to enable horizontal scaling for those services that handle large data loads [68]. In the data pipeline, the services located at the *processing layer* are performing complex and resource-intensive computations that must be performed in a timely manner to maintain the near-real-time character of the system. The rest of the system does not need horizontal scaling, assuming that the system is designed for a private household environment where typically only a few users are tracked. The *processing layer* contains *metric computation* services that produce a unique set of metrics and *user profiling* services that can be used by the *metric computation* services to track and recognize the user of the underlying data at hand. Figure 5.1 illustrates how horizontal scaling can be easily modeled: Assuming there are two microphones streaming audio data to the MQTT topic "audio", if the service *voice-metrics1* cannot keep up with the load of both microphones, the user can let the microphones stream their individual data to their own topics ("audio/mic1", "audio/mic2") and add another *voice-metrics2* service. Then the *voice-metrics1* service could handle all traffic on the topic "audio/mic1" while the *voice-metrics2* service handles the traffic on topic "audio/mic2". In short, horizontal scaling is achieved in this prototype by dividing the data load into fine-grained topics,

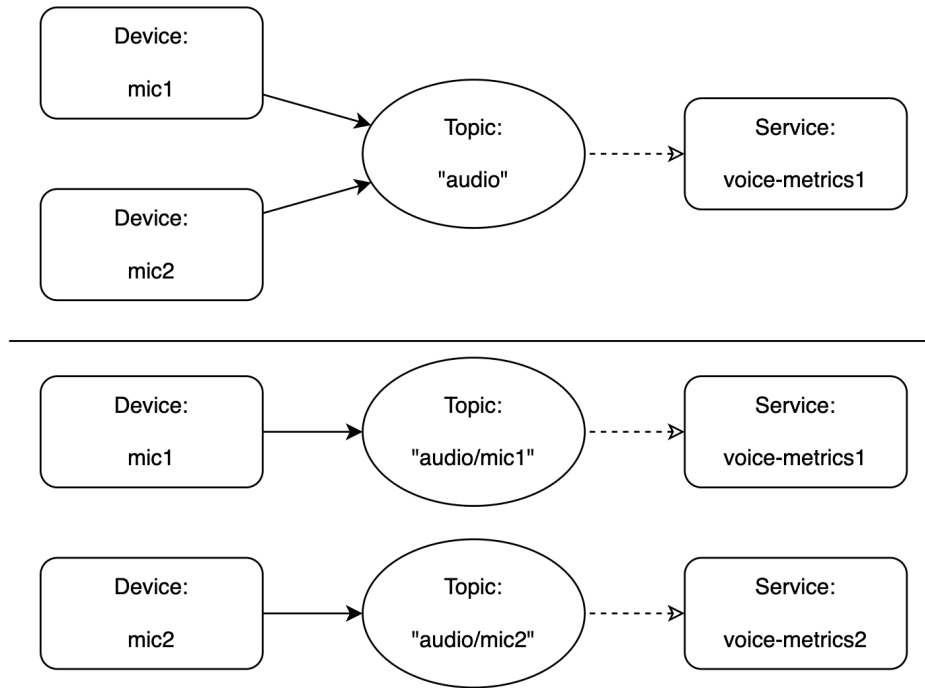


Figure 5.1: Illustration of horizontal scaling on the system's *continuous side*.

allowing individual services to process them independently and distribute the workload efficiently.

The modularity and extensibility of the system are achieved by following the ports and adapters approach for the individual components that are intended for the user to replace or extend, see Table 4.3, *i.e.*, the components *metric computation*, *user profiling*, *temporal context modeling*, and *analysis*. As described in Section 4.3.2.1, the port and adapter principle ensures that the conforming services remain interchangeable and easily integrable [2]. Note again that the implementation for this master's thesis does not strictly conform to the ports and adapters principles in full due to simplicity reasons.

Processing efficiency (NFR-2): The system processes and analyzes real-time sensor data that are continuously collected from the environment. This produces huge amounts of data that need to be processed. If the data input rate exceeds the data processing capacity over a longer period of time, this becomes a classic "bottleneck" situation, where the potential of data loss becomes a risk [119]. To find out whether certain bottlenecks exist in the system, the processing efficiency is measured over the whole architecture under load. For this performance test, processing durations were computed at certain points of the data pipeline. Heavy processing takes place in the first two layers of the data pipeline. Once the *raw_metrics* are persisted in the database, the processing only

continues once the user triggers it. As this part (*i.e.*, the *triggered side*) of the pipeline is not sensitive to performance bottlenecks, processing efficiency is not evaluated from the *persistence layer* onward.

For this performance test, the following files from the Github repository are involved, see Table 5.2 [128]. Note that the worst-case scenario is evaluated: Since the *data ingestion layer* filters for speech data (and thereby reduces the total load on the system), test audio data with continuous speech is used. The duration of the data input segments varies substantially due to the voice activity detection. The used audio file for the performance test consists of concatenated audio samples obtained from the official Toronto Emotional Speech Set (TESS) database [143], see dataset on GitHub [128].

Table 5.2: Involved test files for the performance test [128].

Layer	Filename	Captured information
<i>Data ingestion layer</i>	<i>VoiceFromFilePerformanceTest</i>	Copy of <i>VoiceFromFile</i> that simulates real-time sensing and contains processing duration measures for the steps: <i>collect</i> , <i>filter</i> , <i>transport</i> .
<i>Data ingestion layer</i>	<i>sense_performance_test</i>	Copy of <i>sense.py</i> that starts <i>VoiceFromFilePerformanceTest</i> .
<i>Data ingestion layer</i>	<i>performance_log.csv</i>	Log file for processing duration measures of <i>data ingestion layer</i> .
<i>Processing layer</i>	<i>ComputeMetricsUseCase</i>	Contains processing duration measures for the <i>metrics computation</i> step.
<i>Processing layer</i>	<i>performance_log.csv</i>	Log file for processing duration measures of <i>metrics computation</i> services.
<i>Processing layer</i>	<i>RestUserRecognitionAudioAdapter</i>	Contains processing duration measures for the <i>user profiling</i> step.
<i>Processing layer</i>	<i>performance_log.csv</i>	Log file for processing duration measures of <i>user profiling</i> services.

The audio test file has a total duration of 6 minutes and 40 seconds. When this file was fed into the system, *performance_log.csv*-files were generated, see logfiles on GitHub [128].

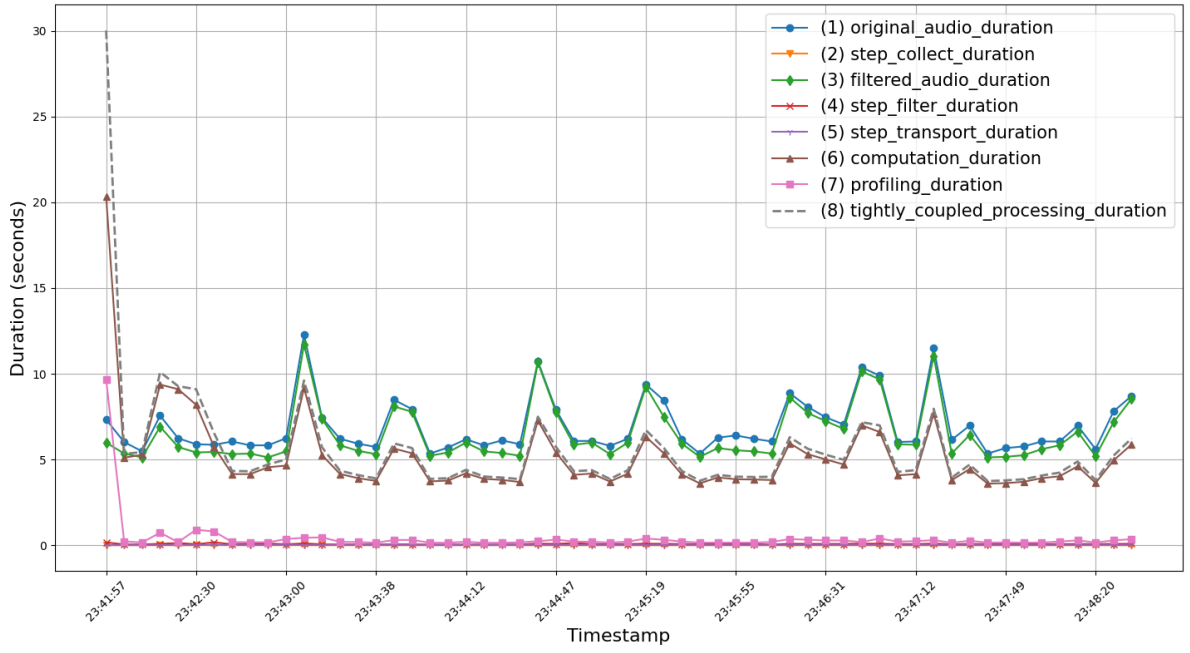


Figure 5.2: Performance test – processing duration measurements.

Figure 5.2 shows the individual processing duration measurements of the test. To make statements about the processing efficiency, it is important to understand the coupling between the services. The metric computation service relies synchronously on the *user profiling* service, and thus the processing times have to be cumulated. The rest of the services involved are loosely coupled. For each of the coupled service groups, the formulated real-time processing constraint, see Algorithm 5.1, must be satisfied.

Algorithm 5.1 Formulated real-time processing constraint.

WHERE:

$t_{\text{processing}}^{(i)}$: Time required to process audio segment i

$t_{\text{audio}}^{(i)}$: Duration of audio segment i

$i \in \text{ProcessingSteps}$: Index of over all tightly coupled service groups

CONSTRAINT:

$$t_{\text{processing}}^{(i)} < t_{\text{audio}}^{(i)} \quad \forall i \in \text{ProcessingSteps}$$

The data shows that the initial data processing times exceed *original_audio_duration* measurements, which indicates a processing bottleneck. However, after the first cou-

ple of segments, the processing capacities are sufficient to maintain the near-real-time characteristic of the system.

Operational simplicity (NFR-3): The initial setup and starting the application are straightforward due to the containerized microservices in Docker compose. User-facing interfaces and visualizations are comprehensible and intuitive. This constitutes a low entry barrier for researchers who are only interested in running their own experiments on the analysis of acoustic voice features. However, with minimal technical knowledge, the proposed system becomes significantly more powerful: Researchers who want to integrate additional modalities are provided with standardized interfaces that enable quick implementations, see Figure 4.4. Note that the corresponding services in the other layers along the data pipeline also need to be added to the system but can be adapted in large parts from the existing services. Professionals in the fields of medicine, psychology, or behavioral science might come up with meaningful instantiations of the *Linkage Framework*. To apply a new concrete mapping, only the *config.json* file needs to be changed. Similarly, changing the initial user baseline can be achieved by modifying the provided *population_baseline.json* file [128], see Figure 5.3.

```

"articulation_rate": { "mean": 4.232, "std": 0.38 },
"f0_avg": { "mean": 148.091, "std": 39.692 },
"f0_range": { "mean": 135.298, "std": 53.529 },
"f0_std": { "mean": 25.808, "std": 10.876 },
"f2_transition_speed": { "mean": 22.096, "std": 3.845 },
"formant_f1_frequencies_mean": { "mean": 1635.678, "std": 55.287 },
"glottal_pulse_rate": { "mean": 104.249, "std": 28.445 },
"hnr_mean": { "mean": -17.393, "std": 11.62 },
"jitter": { "mean": 0.027, "std": 0.006 },
"pause_count": { "mean": 1.285, "std": 0.391 },
"pause_duration": { "mean": 0.954, "std": 0.392 },
"psd-4": { "mean": 3.76e-7, "std": 6.84e-7 },
"psd-5": { "mean": 1.4e-7, "std": 2.03e-7 },
"psd-7": { "mean": 7.5e-8, "std": 1.19e-7 },
"rate_of_speech": { "mean": 3.53, "std": 0.431 },
"rms_energy_range": { "mean": 0.089, "std": 0.064 },
"rms_energy_std": { "mean": 0.016, "std": 0.012 },
"shimmer": { "mean": 0.145, "std": 0.018 },
"snr": { "mean": 3.627, "std": 0.562 },
"spectral_flatness": { "mean": 0.016, "std": 0.005 },
"spectral_modulation": { "mean": 848.361, "std": 214.39 },
"ti13": { "mean": 0.003, "std": 0.003 },
"temporal_modulation": { "mean": 47.424, "std": 10.726 },
"voice_onset_time": { "mean": -14.443, "std": 49.07 },
"voiced16_20": { "mean": 0.018, "std": 0.017 }

```

Figure 5.3: Baseline (*population_baseline.py*) specification file [128].

Explainability (NFR-4): The explainability of the system is rooted in the concrete *MDD Mapping* that is employed in the current system instantiation. One shortcoming of the current implementation is the lack of user-facing information regarding individual feature-to-indicator mappings. While the transparency is visible for the owner of the system, who is able to look at the *config.json* that is currently used by the system, users without the necessary privileges do not see this information. In future adoptions of this system, the user interface could be improved to show the mapping in an interactive form of, *e.g.*, a Sankey diagram, which is a comprehensible visualization method for this purpose. Knowing the mapping and the baseline, together with the resulting DSM-5 *indicator scores* within the system's *dashboard*, the data pipeline is fully transparent, and *indicator score* fluctuations are traceable through the system. An illustration of a complete pass through of simulated $F0_{avg}$ data is depicted in Figure 5.4 and Figure 5.5. The simulated data reflects user behavior divided into three phases: 180 days *baseline behavior*, 180 days *depressed behavior*, and 180 days *recovery behavior*. The figures show the analysis at each individual step: (1) *raw metrics*, (2) *aggregate metrics*, (3) *contextualize metrics*,

(4) *analyze metrics*, and (5) *derive indicator scores*. The corresponding scripts can be found on GitHub [128].

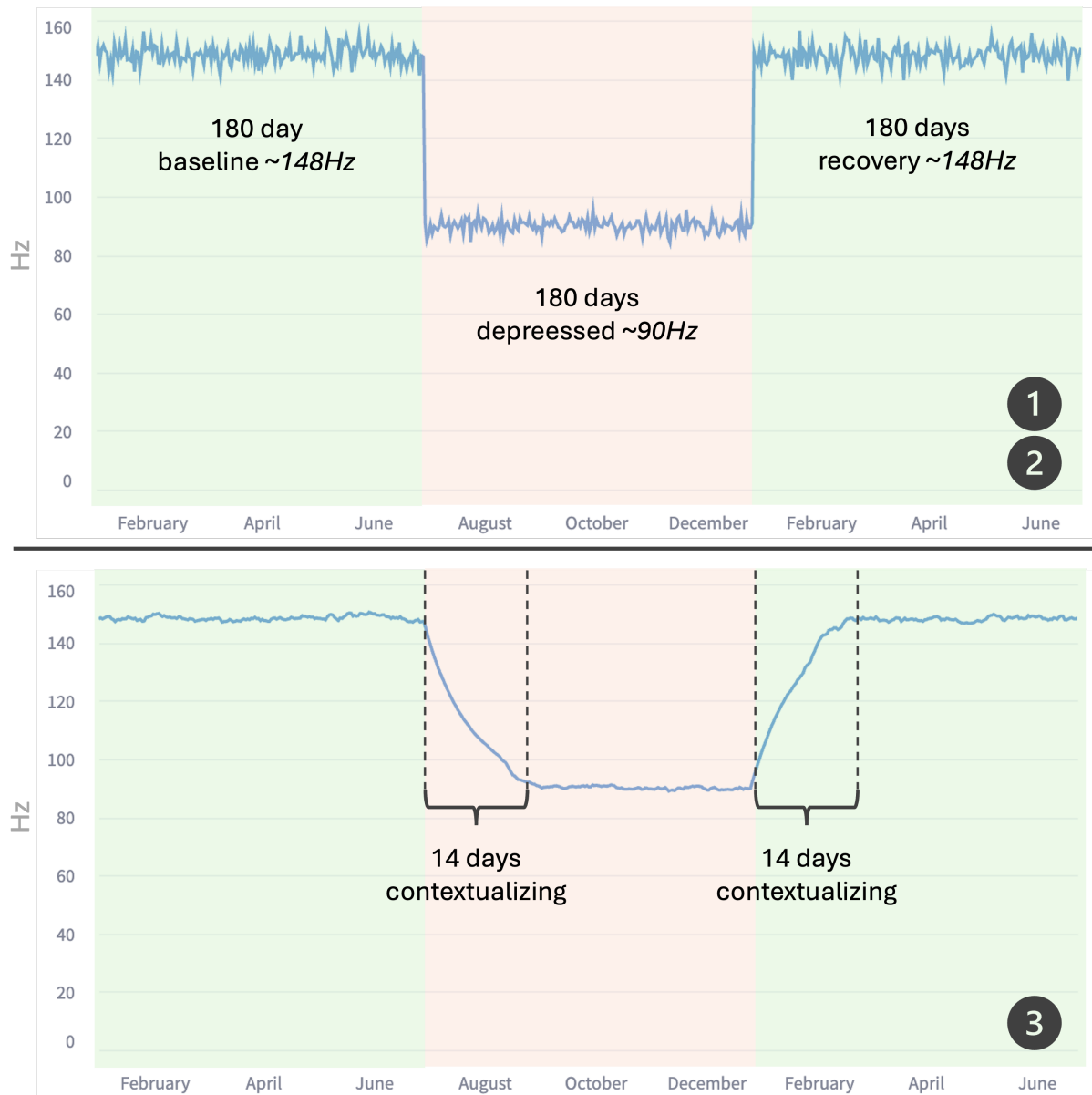


Figure 5.4: Simulated $F0_{avg}$ data – *triggered side* of data pipeline (1/2) [128].

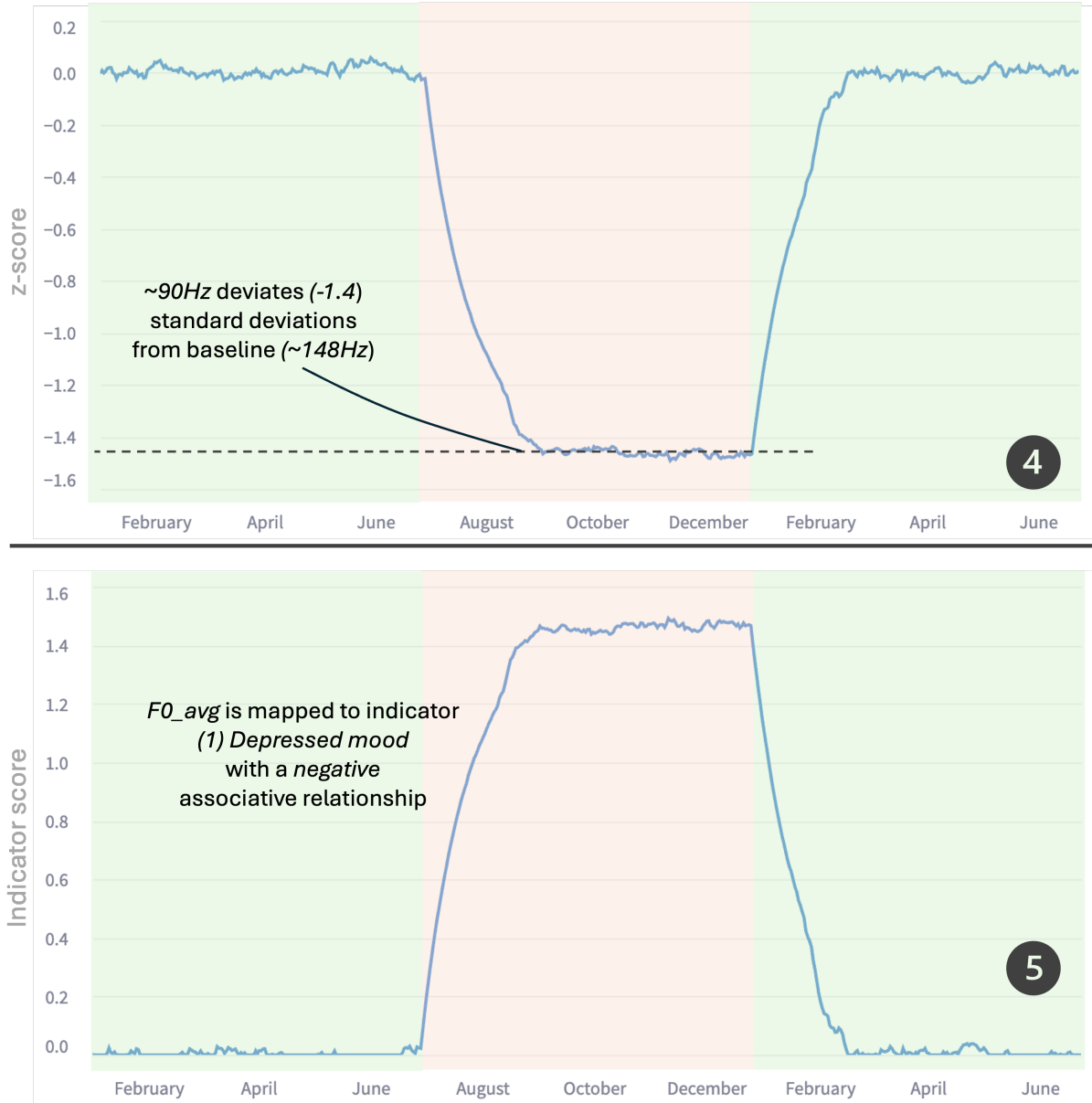


Figure 5.5: Simulated $F0_{avg}$ data – *triggered side* of data pipeline (2/2) [128].

Adaptability (NFR-5): The system has no automated fine-tuning mechanism that adjusts the user’s baseline without human intervention. Unfortunately, the work in this master’s thesis did not produce a novel method to automatically detect user baselines, *i.e.*, the typical ranges of acoustic voice feature values of good mental health specifically for an individual user. Therefore, the adaptability of the system relies on the self-reported mental health status of the user. However, this approach is capable of fine-tuning the user baseline, and the baseline-aware analysis algorithm proposed in Section 4.3.2.7 accounts for changes in the adjusted user baseline, which makes the information provided by the system output even more powerful.

Practical soundness (NFR-6): The practical soundness of the system's outputs depends entirely on the quality of the concrete mapping established by the individual instantiation of the *Linkage Framework*. The evaluation of the concrete *MDD Mapping* can be found in Section 5.2.

5.2 Linkage Framework

The *Linkage Framework* (Obj-2) has been proposed in Section 3.2, together with its concrete instantiation of the *MDD Mapping* in Section 3.3. Figure 5.6 shows the conceptual idea of the *Linkage Framework*, where professionals in the field of mental health care specify the knowledge about any mental health disease, using the unified framework. This builds the foundation for how an automated mental health system could be semantically reflected. This abstraction allows researchers to instantiate their own individual concrete mappings while adhering to the overarching framework to validate their hypotheses. In this way, explainability and interpretability are supported.

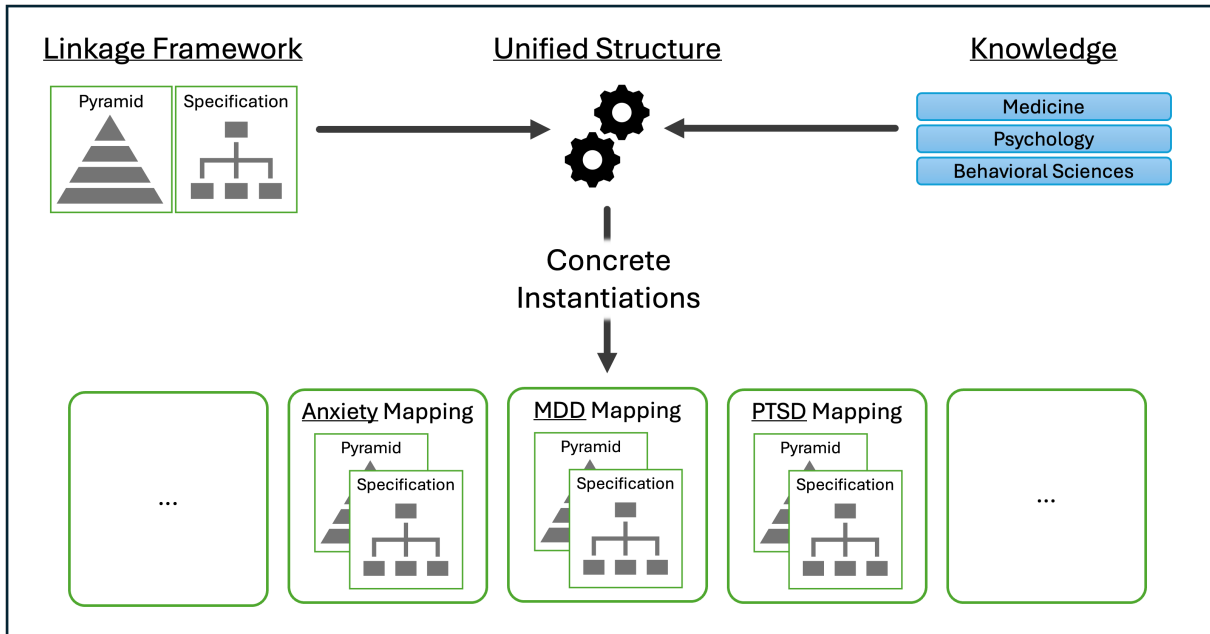


Figure 5.6: Conceptual idea of the *Linkage Framework*.

In the use case of this master's thesis, this structured approach maps low-level features that are directly obtained from edge devices of the IoT-enabled private household environment to the respective high-level DSM-5 indicators [18].

5.2.1 Method

Furthermore, the concrete *MDD Mapping* and its *practical soundness* (NFR-6) are being evaluated. Note that the concrete *MDD Mapping*, see Section 3.3, including the *MDD Mapping Pyramid* and the *MDD Feature Mapping Specification*, is grounded in scientific findings but contains a selective and subjective aspect, which potentially limits its potential and validity.

To show the practical soundness of the mapping, an empirical hypothesis-driven evaluation is conducted, as very few studies explored acoustic voice features towards individual DSM-5 indicators for MDD. The aim is to show that the selected features and their respective correlation are representative of MDD and that they agree with traditional or self-reported assessment methods. Following the hypothetico-deductive (HD) methodology, testable hypotheses are formulated for each individual feature-to-indicator mapping as provided by the *MDD Mapping*, see Table 3.3. These hypotheses are then tested against the ground truth reported in the dataset at hand [225].

5.2.2 Hypotheses

The testable hypotheses formulated are derived from the concrete *MDD Mapping*, see Table 3.3, where each hypothesis reflects one feature-to-indicator mapping. Table 5.3 contains the formulated hypotheses that need to be tested. One-tailed hypotheses are appropriate here [69], as the *Feature Mapping Specification* clearly makes assumptions about the directional relationships between features and indicators. To validate the *MDD Mapping*, the null hypothesis must be rejected. In the case where the null hypothesis can be rejected due to empirical evidence, the alternative hypothesis is accepted, and the particular assumption made by the *MDD Mapping* is correct, *i.e.*, verifying the feature-to-indicator association [225]. Note that in this evaluation the associative relationship, *i.e.*, *gradual*, *nonlinear*, etc., see Section 3.5, is not assessed.

Table 5.3: Formulated hypotheses.

Null hypothesis (H_0)	Alternative hypothesis (H_1)
$H_0 : \rho(F1, \text{Ind}_1) \leq 0$	$H_1 : \rho(F1, \text{Ind}_1) > 0$
$H_0 : \rho(\text{SNR}, \text{Ind}_1) \geq 0$	$H_1 : \rho(\text{SNR}, \text{Ind}_1) < 0$
$H_0 : \rho(F0_{\text{avg}}, \text{Ind}_1) \geq 0$	$H_1 : \rho(F0_{\text{avg}}, \text{Ind}_1) < 0$
$H_0 : \rho(F0_{\text{std}}, \text{Ind}_1) \geq 0$	$H_1 : \rho(F0_{\text{std}}, \text{Ind}_1) < 0$
$H_0 : \rho(\text{SF}, \text{Ind}_1) \geq 0$	$H_1 : \rho(\text{SF}, \text{Ind}_1) < 0$
$H_0 : \rho(\text{Speech rate}, \text{Ind}_1) \geq 0$	$H_1 : \rho(\text{Speech rate}, \text{Ind}_1) < 0$

Null hypothesis (H_0)	Alternative hypothesis (H_1)
$H_0 : \rho(\text{Articulation rate, Ind}_1) \geq 0$	$H_1 : \rho(\text{Articulation rate, Ind}_1) < 0$
$H_0 : \rho(\text{Pause duration, Ind}_1) \leq 0$	$H_1 : \rho(\text{Pause duration, Ind}_1) > 0$
$H_0 : \rho(\text{Pause frequency, Ind}_1) \leq 0$	$H_1 : \rho(\text{Pause frequency, Ind}_1) > 0$
$H_0 : \rho(\text{VSA, Ind}_1) \geq 0$	$H_1 : \rho(\text{VSA, Ind}_1) < 0$
$H_0 : \rho(F2_{\text{transition speed, Ind}_1}) \geq 0$	$H_1 : \rho(F2_{\text{transition speed, Ind}_1}) < 0$
$H_0 : \rho(\text{Jitter, Ind}_1) \leq 0$	$H_1 : \rho(\text{Jitter, Ind}_1) > 0$
$H_0 : \rho(\text{Shimmer, Ind}_1) \leq 0$	$H_1 : \rho(\text{Shimmer, Ind}_1) > 0$
$H_0 : \rho(\text{Volume intensity range, Ind}_2) \geq 0$	$H_1 : \rho(\text{Volume intensity range, Ind}_2) < 0$
$H_0 : \rho(\text{Volume intensity}_{\text{std, Ind}_2}) \geq 0$	$H_1 : \rho(\text{Volume intensity}_{\text{std, Ind}_2}) < 0$
$H_0 : \rho(F0_{\text{range, Ind}_2}) \geq 0$	$H_1 : \rho(F0_{\text{range, Ind}_2}) < 0$
$H_0 : \rho(F0_{\text{std, Ind}_2}) \geq 0$	$H_1 : \rho(F0_{\text{std, Ind}_2}) < 0$
$H_0 : \rho(\text{Jitter, Ind}_2) \leq 0$	$H_1 : \rho(\text{Jitter, Ind}_2) > 0$
$H_0 : \rho(\text{Shimmer, Ind}_2) \leq 0$	$H_1 : \rho(\text{Shimmer, Ind}_2) > 0$
$H_0 : \rho(F0_{\text{avg, Ind}_3}) \geq 0$	$H_1 : \rho(F0_{\text{avg, Ind}_3}) < 0$
$H_0 : \rho(\text{Jitter, Ind}_3) \leq 0$	$H_1 : \rho(\text{Jitter, Ind}_3) > 0$
$H_0 : \rho(\text{Shimmer, Ind}_3) \leq 0$	$H_1 : \rho(\text{Shimmer, Ind}_3) > 0$
$H_0 : \rho(\text{HNR, Ind}_3) \geq 0$	$H_1 : \rho(\text{HNR, Ind}_3) < 0$
$H_0 : \rho(\text{Temporal modulations}_{2-8\text{Hz, Ind}_4}) = 0$	$H_1 : \rho(\text{Temporal modulations}_{2-8\text{Hz, Ind}_4}) \neq 0$
$H_0 : \rho(\text{Spectral modulations}_{2\text{cyc/oct, Ind}_4}) = 0$	$H_1 : \rho(\text{Spectral modulations}_{2\text{cyc/oct, Ind}_4}) \neq 0$
$H_0 : \rho(\text{Speech rate, Ind}_5) \geq 0$	$H_1 : \rho(\text{Speech rate, Ind}_5) < 0$
$H_0 : \rho(\text{Articulation rate, Ind}_5) \geq 0$	$H_1 : \rho(\text{Articulation rate, Ind}_5) < 0$
$H_0 : \rho(\text{Pause duration, Ind}_5) \leq 0$	$H_1 : \rho(\text{Pause duration, Ind}_5) > 0$
$H_0 : \rho(\text{Pause frequency, Ind}_5) \leq 0$	$H_1 : \rho(\text{Pause frequency, Ind}_5) > 0$
$H_0 : \rho(\text{VOT, Ind}_5) \leq 0$	$H_1 : \rho(\text{VOT, Ind}_5) > 0$
$H_0 : \rho(F2_{\text{transition speed, Ind}_5}) \geq 0$	$H_1 : \rho(F2_{\text{transition speed, Ind}_5}) < 0$
$H_0 : \rho(F0_{\text{avg, Ind}_5}) \geq 0$	$H_1 : \rho(F0_{\text{avg, Ind}_5}) < 0$
$H_0 : \rho(F0_{\text{std, Ind}_5}) \geq 0$	$H_1 : \rho(F0_{\text{std, Ind}_5}) < 0$
$H_0 : \rho(\text{SF, Ind}_5) \leq 0$	$H_1 : \rho(\text{SF, Ind}_5) > 0$
$H_0 : \rho(\text{VSA, Ind}_5) \geq 0$	$H_1 : \rho(\text{VSA, Ind}_5) < 0$
$H_0 : \rho(\text{Jitter, Ind}_5) \leq 0$	$H_1 : \rho(\text{Jitter, Ind}_5) > 0$
$H_0 : \rho(\text{Shimmer, Ind}_5) \leq 0$	$H_1 : \rho(\text{Shimmer, Ind}_5) > 0$
$H_0 : \rho(\text{Temporal modulations}_{2-8\text{Hz, Ind}_6}) = 0$	$H_1 : \rho(\text{Temporal modulations}_{2-8\text{Hz, Ind}_6}) \neq 0$
$H_0 : \rho(\text{Spectral modulations}_{2\text{cyc/oct, Ind}_6}) = 0$	$H_1 : \rho(\text{Spectral modulations}_{2\text{cyc/oct, Ind}_6}) \neq 0$
$H_0 : \rho(\text{Jitter, Ind}_8) \leq 0$	$H_1 : \rho(\text{Jitter, Ind}_8) > 0$
$H_0 : \rho(\text{Shimmer, Ind}_8) \leq 0$	$H_1 : \rho(\text{Shimmer, Ind}_8) > 0$
$H_0 : \rho(\text{Pulse (glottal pulses), Ind}_8) \geq 0$	$H_1 : \rho(\text{Pulse (glottal pulses), Ind}_8) < 0$
$H_0 : \rho(\text{Pause duration, Ind}_8) \leq 0$	$H_1 : \rho(\text{Pause duration, Ind}_8) > 0$
$H_0 : \rho(\text{Pause frequency, Ind}_8) \leq 0$	$H_1 : \rho(\text{Pause frequency, Ind}_8) > 0$
$H_0 : \rho(F0_{\text{range, Ind}_8}) \geq 0$	$H_1 : \rho(F0_{\text{range, Ind}_8}) < 0$
$H_0 : \rho(F0_{\text{std, Ind}_8}) \geq 0$	$H_1 : \rho(F0_{\text{std, Ind}_8}) < 0$
$H_0 : \rho(\text{Volume intensity range, Ind}_8) \geq 0$	$H_1 : \rho(\text{Volume intensity range, Ind}_8) < 0$
$H_0 : \rho(\text{Volume intensity}_{\text{std, Ind}_8}) \geq 0$	$H_1 : \rho(\text{Volume intensity}_{\text{std, Ind}_8}) < 0$
$H_0 : \rho(\text{PSD}_4, \text{Ind}_9) = 0$	$H_1 : \rho(\text{PSD}_4, \text{Ind}_9) \neq 0$

Null hypothesis (H_0)	Alternative hypothesis (H_1)
$H_0 : \rho(\text{PSD_5}, \text{Ind}_9) = 0$	$H_1 : \rho(\text{PSD_5}, \text{Ind}_9) \neq 0$
$H_0 : \rho(\text{PSD_7}, \text{Ind}_9) = 0$	$H_1 : \rho(\text{PSD_7}, \text{Ind}_9) \neq 0$
$H_0 : \rho(\text{Voiced-to-silence } (t_{13}), \text{Ind}_9) \leq 0$	$H_1 : \rho(\text{Voiced-to-silence } (t_{13}), \text{Ind}_9) > 0$
$H_0 : \rho(\text{Voiced16:20 PDF}, \text{Ind}_9) \geq 0$	$H_1 : \rho(\text{Voiced16:20 PDF}, \text{Ind}_9) < 0$

5.2.3 DAIC-WOZ Dataset

Since this work formulates the hypotheses on the DSM-5 indicator level, a dataset labeled with corresponding DSM-5 indicator scores is required. One that partially meets this requirement is the Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) database [41]. DAIC-WOZ contains rich multimodal information, such as audio and video data of participants that is collected during clinical interviews [95]. The participants include both distressed and non-distressed individuals with different diseases such as anxiety, depression, post-traumatic stress disorder (PTSD), etc. [95]. The dataset is part of a larger effort to identify non-verbal and verbal biomarkers of various mental disorders [45]. Its popularity is undeniable [226]: In a study by Mao *et al.*, it was shown that 39% of the previous work on depression detection utilized this DAIC-WOZ dataset [112].

Furthermore, the dataset contains PHQ-8 questionnaire scores (similar to PHQ-9) that are used for depression assessments [110]. The *MDD Mapping* relies on the PHQ-9 questionnaire that directly links its items to the DSM-5 clinical indicators. However, a direct link to PHQ-8 is possible. Both questionnaires are identical, except that PHQ-8 removes the ninth item (*suicidality*) due to its questionable contribution to correctly assessing depression in general. Ultimately, the equivalency has been investigated by multiple parties [36, 83, 242], leading to the conclusion that the PHQ-8 questionnaire is equally reliable as the PHQ-9 questionnaire [36]. Therefore, note that in this evaluation, the work only focuses on the first eight items on the PHQ-8/-9 questionnaire. This enables direct testing of the DSM-5 related indicators with their corresponding acoustic voice features. See Table 5.4 for the comparison of PHQ-8, PHQ-9, and the MDD DSM-5 clinical indicators and how they are linked for this evaluation.

Table 5.4: PHQ-8, PHQ-9 questionnaire items [109, 110] and the MDD DSM-5 clinical indicators [18].

PHQ-8 item	PHQ-9 item	DSM-5 indicator
1. Little interest or pleasure in doing things	1. Little interest or pleasure in doing things	(2) Loss of interest
2. Feeling down, depressed, or hopeless	2. Feeling down, depressed, or hopeless	(1) Depressed mood
3. Trouble falling or staying asleep, or sleeping too much	3. Trouble falling or staying asleep, or sleeping too much	(4) Insomnia / hypersomnia
4. Feeling tired or having little energy	4. Feeling tired or having little energy	(6) Loss of energy
5. Poor appetite or overeating	5. Poor appetite or overeating	(3) Significant weight changes
6. Feeling bad about yourself, or that you are a failure, or have let yourself or your family down	6. Feeling bad about yourself – or that you are a failure or have let yourself or your family down	(7) Worthlessness / guilt
7. Trouble concentrating on things, such as reading the newspaper or watching television	7. Trouble concentrating on things, such as reading the newspaper or watching television	(8) Diminished ability to think / concentrate
8. Moving or speaking so slowly that other people could have noticed. Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	8. Moving or speaking so slowly that other people could have noticed. Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	(5) Psychomotor retardation / agitation
—	9. Thoughts that you would be better off dead, or of hurting yourself	(9) Suicidality

5.2.4 Hypotheses Testing

From the DAIC-WOZ dataset, a random selection of 64 participants ($n = 64$) was chosen while maintaining a balanced gender distribution (50% male, 50% female). Voice characteristics differ substantially between male and female individuals [144]. For each

participant, the *Participant_ID*, *gender*, and all eight self-reported PHQ-8 indicator scores were collected. Furthermore, the audio data was fed through the system to extract the selected acoustic voice features. Each participant's audio data was segmented by the system, resulting in between 300 and over 2,000 individual *raw_metric* values per participant. To simplify the evaluation, the mean value for each extracted metric was computed across all segments per participant.

To protect participant confidentiality, the original *Participant_ID* has been reassigned with a new *Participant_ID_Disguised* value. The final dataset, aggregated into the *evaluation_dataset.csv* file, is available in the GitHub repository [128].

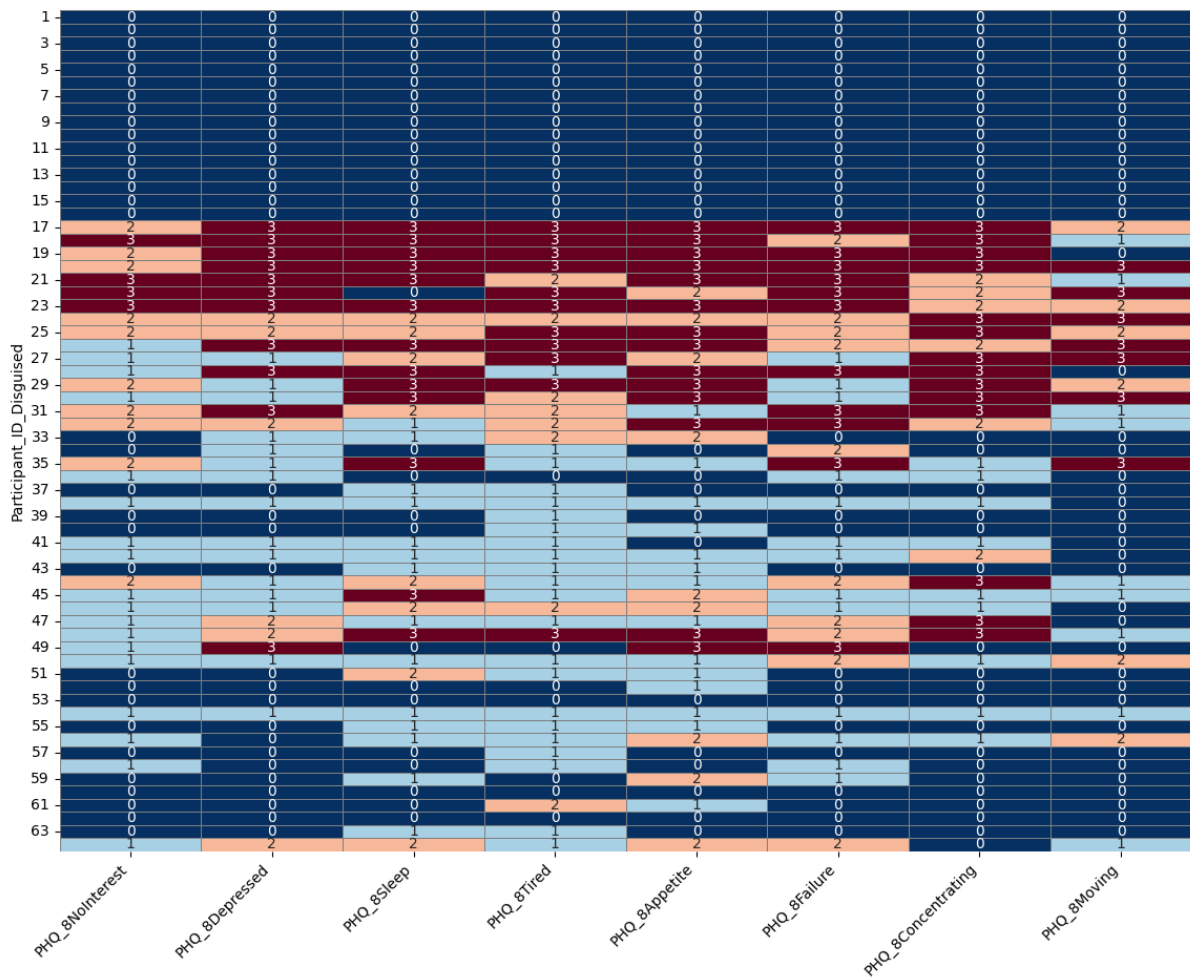


Figure 5.7: Self-reported PHQ-8 indicator scores across participants.

Figure 5.7 displays the distribution of the self-reported PHQ-8 indicator scores. The Jupyter notebook for this upcoming analysis, *evaluation_mdd_mapping.ipynb*, is also

available on GitHub [128]. The complete set of additional resources of this analysis is collected in the Appendix C.

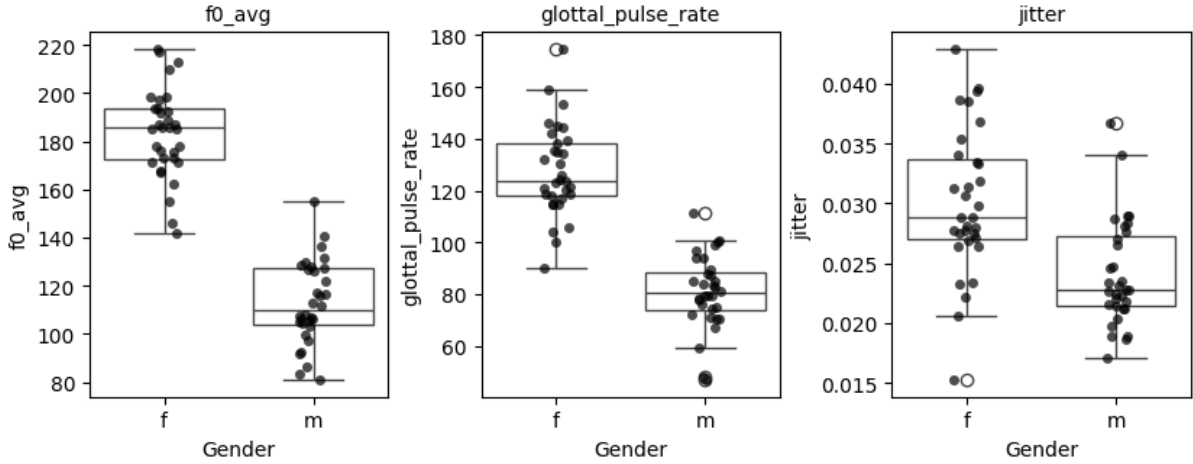


Figure 5.8: Distribution of selected male and female acoustic voice features.

By plotting the participant data by the collected voice features, see Figure 5.8, it is revealed that some of the features have differences between male and female participants, such as *f0_avg*, *glottal_pulse_rate*, or *jitter*. Figure 5.9 shows the differences of each voice feature, quantified by the *Cohen's d* value. Typically, if the *Cohen's d* value is below $\leq \pm 0.2$, the difference between two groups is considered negligible [170]. To avoid testing hypotheses where strong gender-specific effects are present, the dataset is divided into male and female groups where at least a small effect is observed.

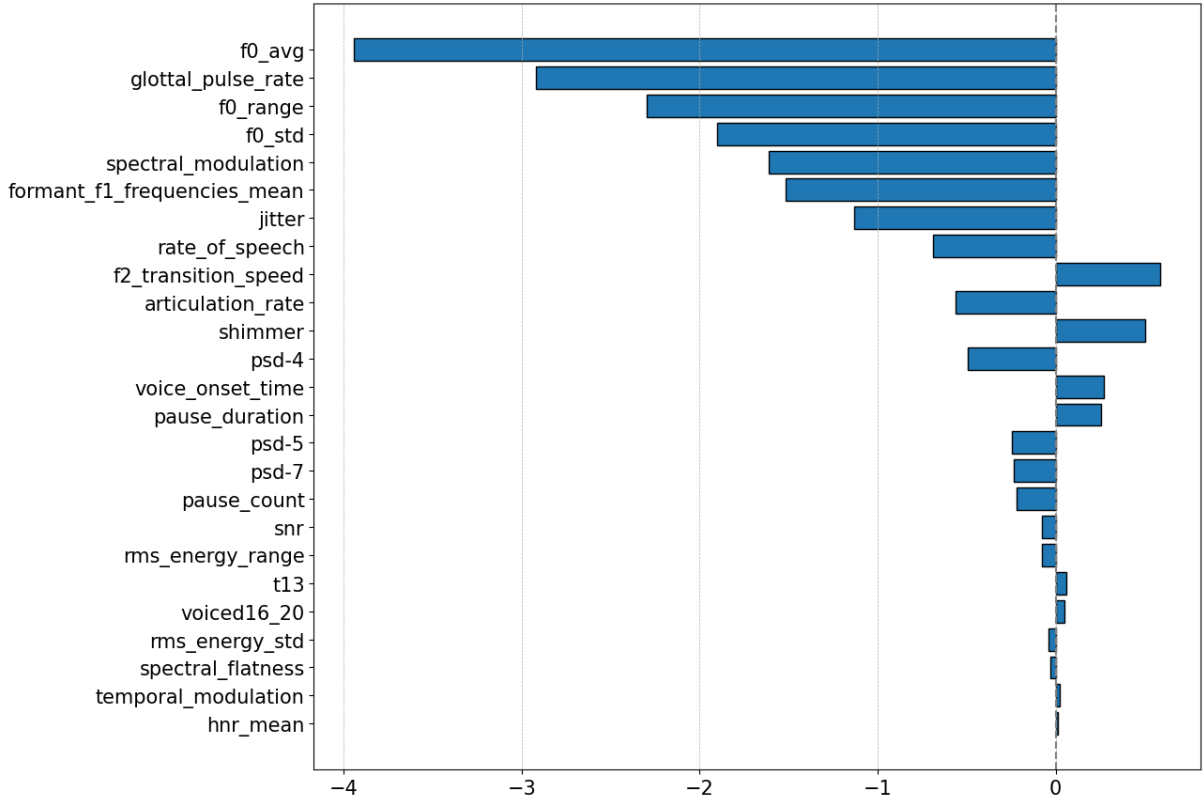


Figure 5.9: *Cohen's d* (positive = higher in males, negative = higher in females).

To evaluate the defined hypotheses, the Pearson Correlation Coefficients (PCC) were computed for each pair of acoustic voice feature and MDD DSM-5 indicator [40]. This analysis tries to identify potentially relevant feature-to-indicator mappings that are not explicitly proposed in the original *MDD Mapping*.

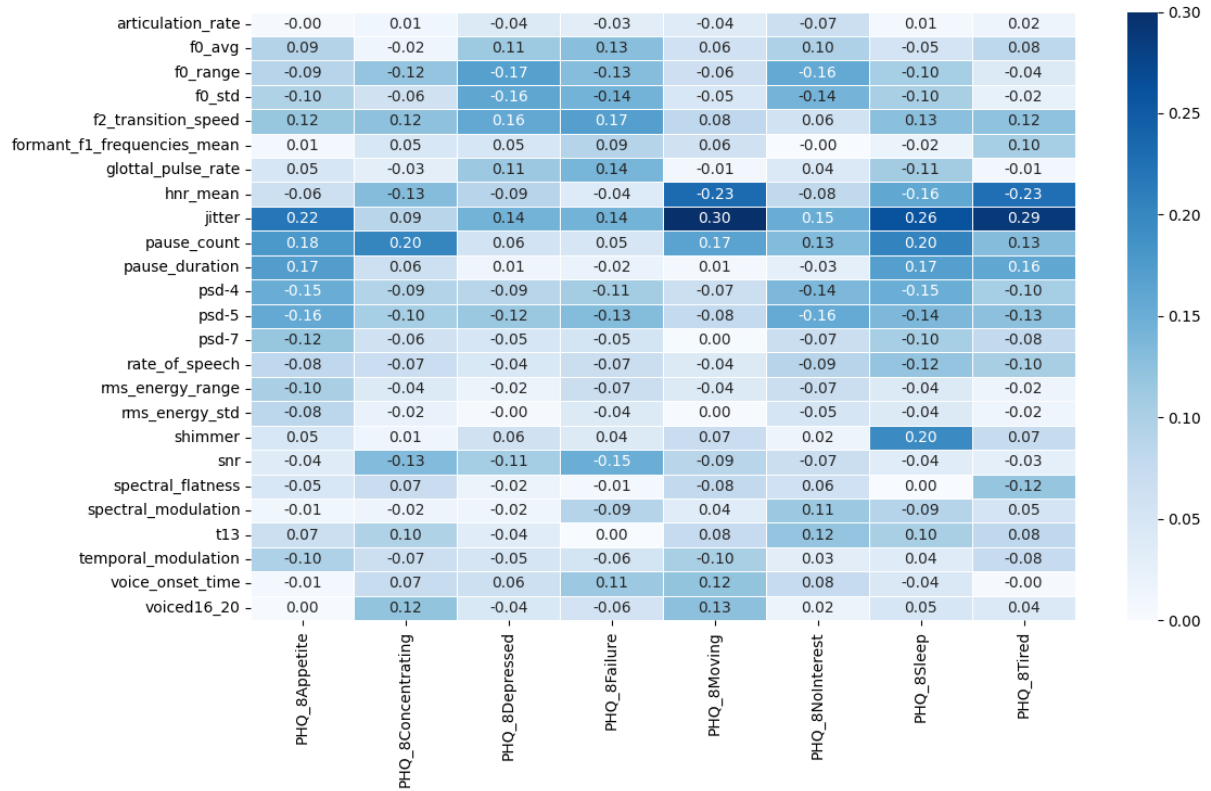


Figure 5.10: Pearson correlation coefficients of feature-to-indicator pairings.

Figure 5.10 shows the effect sizes (r correlation coefficients), and Figure 5.11 shows the respective statistical significance of the observed relationship (p-value) [40]. Optimal candidates exhibit high effect sizes with statistical significance.

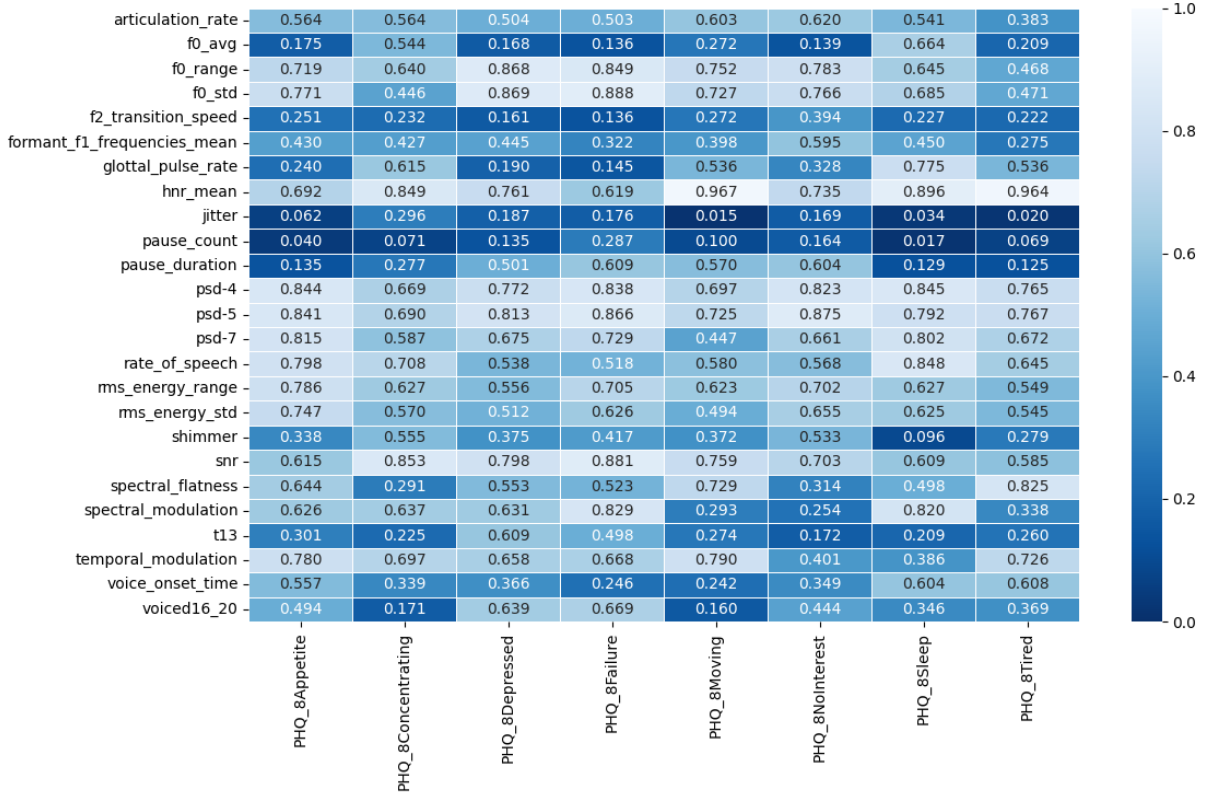


Figure 5.11: Pearson p-values of feature-to-indicator pairings.

Although several feature-to-indicator pairs show statistically significant correlations, the large number of comparisons increases the risk of *Type I* errors (false positives) [100]. Therefore, false discovery rate (FDR) adjustments were applied, using the Benjamini-Hochberg method [156], resulting in drastically deteriorated p-values, see Appendix C. Assuming a well-established significance threshold of $p = 0.05$, none of the correlations is statistically significant: None of the null hypotheses can be rejected. The evidence shows that the proposed concrete *MDD Mapping* does not fulfill the non-functional requirement *practical soundness* (*NFR-6*). These results and potential explanations are further discussed in Section 5.3.

5.2.5 Second Attempt

To evaluate the system's integration of the *Linkage Framework* and its *practical soundness* (*NFR-6*), a second test dataset is utilized, the Toronto Emotional Speech Set (TESS) [143]. For this test, all available speech samples for one individual participant have been concatenated such that two *wav* files are created: the first contains only *non-depressed* speech (*long_nondepressed_sample_nobreak.wav*), while the second contains *depressed* speech (*long_depressed_sample_nobreak.wav*) [128]. Both files were simulated and fed

into the system as if the samples were recorded on consecutive days. After triggering and running the analysis of the data, the following DSM-5 *indicator scores* are reported over time, see Figure 5.12.

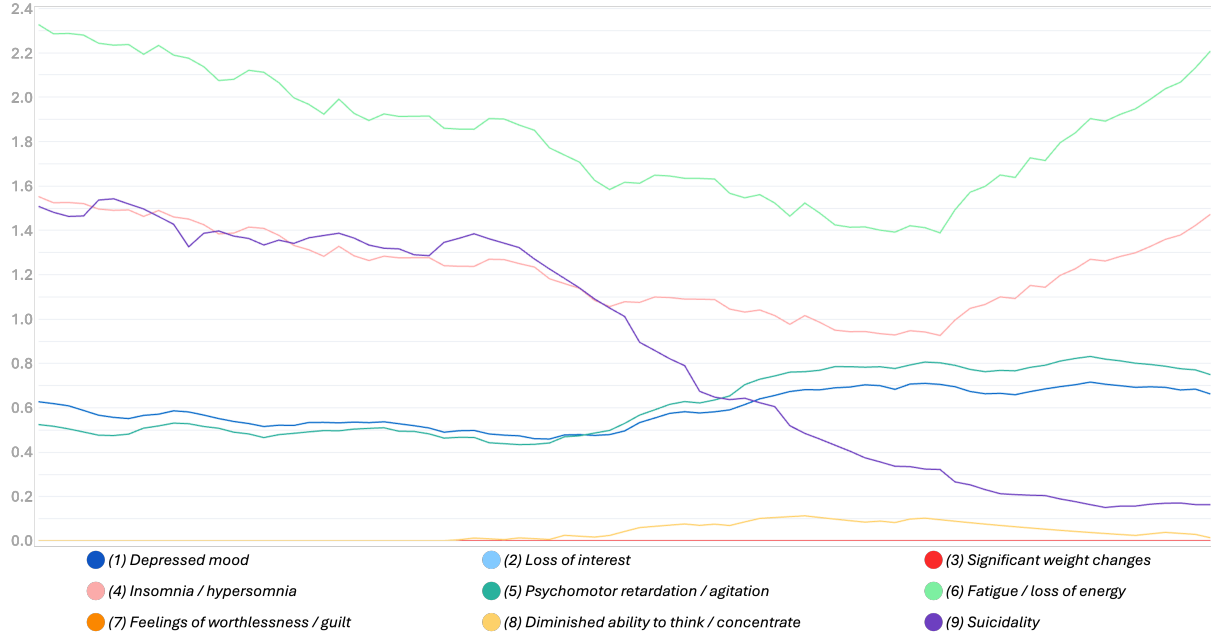


Figure 5.12: DSM-5 *indicator scores* for simulated streamed data of TESS dataset.

This shows that the system gradually detects depressive speech patterns, resulting in higher MDD DSM-5 *indicator scores*. Since each indicator in the formulated hypotheses contributes individually to the overall assessment of depression, observing whether individual acoustic voice features correlate with the aggregated depression trend provides a weak validation of the *MDD Mapping*. Table 5.5 shows the results at the feature-to-indicator level, highlighting whether the concrete *MDD Mapping* supports the overarching hypothesis of acoustic voice features being indicative of depressive behavior.

Table 5.5: Evaluation of directional hypotheses on acoustic voice features.

Metric	Dir.	Depressed mean	Non-depressed mean	Confirmed
formant_f1_frequencies_mean	>	1655.9247	1719.4960	×
snr	<	3.8749	3.9476	✓
f0_avg	<	197.3006	267.1643	✓
f0_std	<	41.6216	97.6998	✓
spectral_flatness	<	0.0141	0.0219	✓
rate_of_speech	<	2.5122	3.4348	✓
articulation_rate	<	2.9756	3.7826	✓
pause_duration	>	0.7512	0.5565	✓
pause_count	>	1.7317	1.4348	✓
f2_transition_speed	<	15.4186	19.2135	✓

Metric	Dir.	Depressed mean	Non-depressed mean	Confirmed
jitter	>	0.0117	0.0248	×
shimmer	>	0.0606	0.0814	×
rms_energy_range	<	0.1723	0.2139	✓
rms_energy_std	<	0.0343	0.0432	✓
f0_range	<	218.7375	388.5426	✓
hnr_mean	<	4.4823	-0.9138	×
temporal_modulation	≠	34.6273	65.6325	✓
spectral_modulation	≠	1568.0581	1220.3354	✓
voice_onset_time	>	16.8261	30.3561	×
glottal_pulse_rate	<	150.7611	198.2620	✓
psd-4	≠	0.0000	>0.0000	✓
psd-5	≠	0.0000	>0.0000	✓
psd-7	≠	0.0000	>0.0000	✓
t13	>	0.0043	0.0041	✓
voiced16_20	<	0.0356	0.0071	×

5.3 Discussion

The evaluation of the *System Architecture* shows that 23 FRs have been *fully implemented*, 8 FRs are *partly implemented*, and 3 FRs have been *intentionally rejected* due to lack of necessity. No FRs are *not implemented* at the current development stage, see Table 5.6.

Table 5.6: Overview of functional requirement (FR) evaluation.

Status	Description	Count
✓	Fully implemented	23
(✓)	Partly implemented	8
–	Not implemented	0
×	Rejected	3

For the non-functional requirement evaluation, the *processing efficiency (NFR-2)* was interesting in particular. It greatly depends on the proportion between the data input rate and the processing capacity of the *processing layer* services. For instance, an unfiltered video stream creates a lot of input data. If a dedicated *metrics computation* service performs complex calculations on these data, the processing capacity of the *metrics computation* service might be below the data input rate. Therefore, smart decisions must be made regarding data preprocessing in the *data ingestion layer*, as well as whether to decompose the holistic *metrics computation* service into multiple smaller services, each responsible for a subset of metrics. The user of the system has to understand the implications of different modalities and their effect on the processing efficiency.

These results, paired with the positive assessment of the non-functional requirements, show that the proposed *System Architecture* is a robust prototypical system as an attempt for a generalizable automated mental health assessment system that may benefit from future adoptions. It would be interesting to see different domain areas for this blueprint concept, where different modalities and computation techniques are applied along the data pipeline.

The evaluation of the concrete *MDD Mapping* has unfortunately not confirmed the *practical soundness (NFR-6)* of the proposed concrete *MDD Mapping*. None of the formulated hypotheses for the individual mappings were statistically significant. In fact, the reasons for this are manifold: First, a key limitation lies in the size of the dataset, with only 64 participants, showing no realistic distribution of the general population. In addition, self-reported PHQ-8 assessments of the dataset introduce subjectivity, making the detection of unbiased correlations even more difficult. However, none of the null hypotheses could be rejected; individual acoustic voice features showed a certain degree of effect, such as *jitter* or *pause_count*. It would be interesting to conduct a larger study and evaluate the concrete *MDD Mapping* with more data, the more complete PHQ-9 questionnaire, and potentially even longitudinal data, capturing participant's voice data and their mental health states over a period of multiple months or even years. This would allow conclusions to be drawn between the MDD DSM-5 *indicator scores* and the respective acoustic voice features of an individual participant, thus reducing the impact of the highly individual nature of the human voice. However, this study demonstrated that acoustic voice features must be considered individually. However, if the overarching assumption that depression patterns are generalizable in the context of acoustic voice features does not hold, then the utility of the *Linkage Framework* would be entirely questionable.

As a second approach to demonstrating the *practical soundness (NFR-6)* of the concrete *MDD Mapping*, a holistic view of the system reveals the clear benefit of incorporating a personalized baseline. By allowing users to provide self-reported PHQ-9 responses, the system can adjust baseline parameters in a user-specific manner and therefore produce even more powerful insights into the state of mental health. By testing the system with speech data from another public speech dataset (TESS) and testing its outputs against the labeled data, it has been shown that the concrete *MDD Mapping* is indeed able to indicate meaningful tendencies in the MDD DSM-5 *indicator scores*. For this final step of the evaluation, it is absolutely necessary to point out that the system is a

prototypical implementation, including various complex processing steps, implemented in an accessible manner for future research endeavors. Therefore, implementations do not meet industry-grade standards, highlighting the need for future work.

Despite the results, there is good news: The *Linkage Framework* precisely fulfills its job as intended. It is now in the hands of future research to come up with more practically sound concrete mappings. By representing a bridge between feature extraction and hypothesis specification, users can easily modify the *config.json* file to define new mappings, allowing rapid hypothesis testing and scientific iteration.

Chapter 6

Considerations & Future Work

6.1 Summary

This master's thesis investigated a novel approach for automated mental health assessment focusing on acoustic voice features to make computational reasoning on mental health states explainable and traceable. In the context of an IoT environment, this thesis proposes a concept, framework, and implementation for the useful and efficient integration of multimodal data streams with near-real-time character to promote scientific collaboration and innovation. While providing a generalizable strategy for automated mental health assessments, this master's thesis has a clear focus on the detection of MDD leveraging the acoustic information contained in the human's voice.

The formulated research objectives (*Obj-1 – Obj-4*), see Section 1.3, inspired the individual scientific contributions of this thesis. After performing an extensive literature review, a generalizable solution to the complex problem of depression detection has been proposed, namely the *Linkage Framework*, consisting of the two conceptual artifacts the *Mapping Pyramid* and the *Feature Mapping Specification*. Both artifacts are needed to provide an initial basis for medical professionals to define and hypothesize about the relationship between physiological and behavioral biomarkers and the corresponding mental health disorder. This framework bridges the semantic gap between the abstract medical diagnosis via DSM-5 clinical indicators (explainable and traceable) and the concrete, measurable world to enable computational reasoning on mental health states. Thereby, the *Linkage Framework* is used as a "blueprint" scheme that unifies all possible mapping instantiations for mental health disorders.

In Chapter 4, a *System Architecture* was presented that can consume such a *Linkage Framework* blueprint instantiation and thereby adopt the semantic reasoning of it. This proposed *System Architecture* is a platform for multimodal data ingestion, data

processing, and data analysis, and together in combination with meaningful and practical sound *Linkage Framework* instantiations, it enables passive, non-intrusive, efficient, and effective automated mental health assessments.

To showcase an end-to-end prototype of this system for acoustic-based voice depression detection, Section 3.4 investigates scientifically grounded acoustic voice features for the individual DSM-5 clinical indicators for MDD. This concrete *MDD Mapping* (instantiation of *Linkage Framework*) has been proposed and plugged into the concrete *System Architecture* implementation of this master’s thesis, see the repository on GitHub [128]. Furthermore, this applied research project was evaluated along pre-defined functional and non-functional requirements that aim to answer the question of whether the two proposed artifacts are truly able to address the research question with its defined research objectives.

The prototype served as a platform for multimodal (auditory, imagery, or textual) data ingestion, data processing, and data analysis, building the foundation for further refinement of the system to leverage knowledge from diverse areas, further hypothesis testing of newly acquired medical insights, and advances in sensing techniques and sensor feature extraction methods of any input modality.

6.2 Considerations

The development of this system revealed the deeper value of an explainable automated mental health assessment method in enabling users to understand how the system comes to its conclusions. This not only builds trust in the system but also holds potential for supporting treatment decisions. A relevant factor was to model the system from two ends: transforming knowledge about mental health disorder into a machine-interpretable format (*Linkage Framework*) and modeling a system as a data pipeline reflecting individual processing steps to reflect the layers of the *Mapping Pyramid*. The contract for how the instantiated *Linkage Framework* is integrated and employed by the *System Architecture* is a key contribution of this master’s thesis and enables the independent, interdisciplinary research and development on both the linkage and the system.

Working on such systems is not only technically challenging but also requires interest and understanding in many areas of computer science. Identifying the core steps of the data pipeline was beneficial in decomposing and decoupling complexities in early iterations. With the help of the artifacts proposed in this work, the entry barriers and challenges faced when creating such a system from scratch can be reduced.

In the context of IoT sensing environments, the need for a multimodal approach has been identified. The focus on acoustic voice features is a prime example highlighting the effort required to integrate any single modality into the system. Tailored for each individual sensor type and modality, data must be collected, features must be extracted, and methods to identify the user must be implemented. These are the steps on the *continuous side* of the system that are not generalizable, where the analytical part (*triggered side*) of the system provides a common solution for any sensor type and modality.

The evaluation of the concrete *MDD Mapping* using the hypothetico-deductive methodology showed no statistical significance for its individual feature-to-indicator mappings. There are many potential explanations for this. Due to the limited selection of the DAIC-WOZ dataset, the sample data do not reflect the voice characteristics of the general population. In addition, the self-reported PHQ-8 assessments introduce subjectivity, and the missing longitudinal data does not support an individual evaluation of the mapping. From this evaluation, the most promising acoustic voice features were *jitter* and *pause_count*, see Section 5.2.4. However, using the TESS dataset, the system outputs that were produced by employing the concrete *MDD Mapping* agreed with the labeled data in 76% of the individual feature-to-indicator mapping items, see Section 5.2.5. These results align with previous literature where studies cannot find consensus on whether a reliable association between acoustic voice features and depression detection can be made. The findings of the evaluation suggest that voice characteristics must be evaluated individually for each user and that feature-to-indicator mappings could be user dependent.

A key takeaway is the notion of a user baseline that can be adjusted through user feedback via the PHQ-9 questionnaire. This baseline is used for the comparison of fresh data and the user's typical behavior to compute indicator scores. However, its capabilities are limited due to the challenge of correctly defining and detecting typical user behavior and the degree to which the baseline is adjusted because of non-linear associative relationships between features and indicators, see Section 3.5.5.

While evaluating the processing efficiency of the individual services, it was shown that extracting the features is the computationally most expensive step. Currently, the system relies on third-party libraries such as *OpenSMILE*, *Praat(-parselmouth)*, and *Librosa*, which limits opportunities for performance optimization. Processing efficiency in particular needs to be monitored with caution to maintain the near-real-time character constraint of the system as described in Section 5.1.2. The processing capacity of

tightly coupled service groups must never exceed the original rate at which data is ingested. Future system adoptions should focus on making smart choices about how edge processing can be performed, based on the specific modality and sensor type at hand, *e.g.*, detect voiced audio segments using voice activity detection.

During the development process, certain limitations have been identified. Different modalities need their own *user profiling* services. In the scenario where more than one user is monitored with multiple modalities, the system is not able to identify the correct user for the computed feature data, *e.g.*, in an instantiation that uses acoustic voice features along with facial recognition features, the system would not be able to associate the features across the two modalities for a certain user.

6.3 Future Work

The software prototype itself builds the foundation for potential future work. Further research in biomarker exploration is needed to find the most promising biomarkers to improve the *practical soundness* of the concrete *MDD Mapping*. A collaboration between experts from diverse fields would be beneficial. As the mapping decomposes depression into the DSM-5 indicators, relationships between indicators and biomarkers need to be identified. Furthermore, biomarker exploration in general becomes more important to support knowledge transfer and advance the field of automated mental health assessment methods to expand the medical knowledge upon which concrete *Linkage Framework* instantiations can be constructed.

For now it was assumed that the associative relationships between features and indicators were gradual, see Section 3.5.5. Choosing better techniques for modeling these relationships (especially non-linear relationships) and strategies for quantifying the feature-to-indicator mappings would improve the effectiveness of the system.

A driving factor for the development of such systems is the availability of datasets. While most existing corpora focus only on a binary depression classification, this novel approach relies on information on the more fine-grained DSM-5 indicators. Therefore, datasets must be annotated at the DSM-5 indicator level. Especially audio datasets are scarce, as many of them rely on actors that simulate depressed speech. Future work should focus on collecting audio datasets labeled on the DSM-5 indicator level with real participants.

In addition, investigating changes in the user's behavior is an integral component of the proposed system. Therefore, collecting user data that spans multiple months or years

would support and enhance the understanding of defining and identifying a user's baseline. Future efforts could explore new strategies for automatically identifying user baselines, as the implemented feedback loop mechanism based on the PHQ-9 questionnaire violates the automated idea.

While each modality employs its own strategy for identifying users, the system currently is not yet capable of associating users across different modalities. Future efforts should explore new methods that enable cross-modal user identification.

Further research may also include refining the functional and non-functional requirements of the system, aligning them better with the purpose of the conceptual idea for a generalizable framework. This implementation foresees but lacks the concrete integration of multiple modalities. Therefore, it would be interesting to see different adoptions of the conceptual artifacts presented in this master's thesis that touch other application areas with different modalities, feature extraction, user profiling, and temporal context modeling techniques for further evaluation and to better align the *Linkage Framework* and the *System Architecture* with the broader objective of advancing automated mental health assessment methods.

Bibliography

- [1] A. C. Trevino, T. F. Quatieri, and N. Malyska. “Phonologically-based biomarkers for major depressive disorder”. In: *EURASIP Journal on Advances in Signal Processing* 2011 (2011), p. 42. DOI: [10.1186/1687-6180-2011-42](https://doi.org/10.1186/1687-6180-2011-42).
- [2] A. Cockburn. *The hexagonal (ports & adapters) architecture*. Version 0.9, to be updated after reader comments. Sept. 2005. URL: <https://alistair.cockburn.us/hexagonal-architecture>.
- [3] A. G. M. Meque, N. Hussain, G. Sidorov, et al. “Machine learning-based guilt detection in text”. In: *Scientific Reports* 13 (2023), p. 11441. DOI: [10.1038/s41598-023-38171-0](https://doi.org/10.1038/s41598-023-38171-0).
- [4] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton. “Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression”. In: *Journal of Psychiatric Research* 27.3 (1993), pp. 309–319. ISSN: 0022-3956. DOI: [https://doi.org/10.1016/0022-3956\(93\)90041-Y](https://doi.org/10.1016/0022-3956(93)90041-Y).
- [5] A. J. Rush, H. A. Sackeim, C. R. Conway, M. T. Bunker, S. D. Hollon, K. Demyttenaere, A. H. Young, S. T. Aaronson, M. Dibué, M. E. Thase, et al. “Clinical research challenges posed by difficult-to-treat depression”. In: *Psychological Medicine* 52.3 (2022), pp. 419–432. DOI: [10.1017/S0033291721004943](https://doi.org/10.1017/S0033291721004943).
- [6] A. Koster, C. H. van Gool, G. I. Kempen, B. W. Penninx, J. S. Lee, S. M. Rubin, F. A. Tylavsky, K. Yaffe, A. B. Newman, T. B. Harris, M. Pahor, H. N. Ayonayon, J. T. van Eijk, S. B. Kritchevsky, and Health ABC Study. “Late-life depressed mood and weight change contribute to the risk of each other”. In: *The American Journal of Geriatric Psychiatry* 18.3 (2010), pp. 236–244. DOI: [10.1097/JGP.0b013e3181c65837](https://doi.org/10.1097/JGP.0b013e3181c65837).
- [7] A. L. Sangiovanni-Vincentelli, L. P. Carloni, F. De Bernardinis, and M. Sgroi. “Benefits and challenges of platform-based design”. In: *Proceedings of the 41st Annual Design Automation Conference (DAC)*. 2004, pp. 409–414.
- [8] A. M. Werner, M. Ernst, E. Brähler, A. N. Tibubos, D. Otten, I. Reiner, J. Wiltink, M. Michal, A. Schulz, P. S. Wild, T. Münzel, J. König, K. J. Lackner, N. Pfeiffer, and M. E. Beutel. “The association of depressive symptoms and body weight change in midlife – Results from the Gutenberg Health Study in Germany”. In: *Journal of Affective Disorders* 332 (2023), pp. 115–124. ISSN: 0165-0327. DOI: <https://doi.org/10.1016/j.jad.2023.03.078>.
- [9] A. Mainous, V. Wu, L. Yin, P. Sharma, B. Jenkins, A. Saguil, D. Nelson, and F. Orlando. “Body mass index versus percent body fat as a predictor of mortality in men aged 20–49”. In: *The Annals of Family Medicine* 22.Supplement 1 (2024). ISSN: 1544-1709. DOI: [10.1370/afm.22.s1.6079](https://doi.org/10.1370/afm.22.s1.6079).
- [10] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes. “Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk”. In: *IEEE Transactions on Biomedical Engineering* 51.9 (2004), pp. 1530–1540. DOI: [10.1109/TBME.2004.827544](https://doi.org/10.1109/TBME.2004.827544).

- [11] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, D. M. Wilkes. “Analysis of fundamental frequency for near term suicidal risk assessment”. In: *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics. 'cybernetics evolving to systems, humans, organizations, and their complex interactions' (cat. no.0. Vol. 3. 2000, 1853–1858 vol.3. DOI: [10.1109/ICSMC.2000.886379](https://doi.org/10.1109/ICSMC.2000.886379).*
- [12] A. Sharma and U. Ghose. “Sentimental analysis of twitter data with respect to general elections in India”. In: *Procedia Computer Science* 173 (2020), pp. 325–334. DOI: [10.1016/j.procs.2020.06.038](https://doi.org/10.1016/j.procs.2020.06.038).
- [13] A. T. Behan, C. Byrne, M. J. Dunn, G. Cagney, and D. R. Cotter. “Proteomic analysis of membrane microdomain-associated proteins in the dorsolateral prefrontal cortex in schizophrenia and bipolar disorder reveals alterations in LAMP, STXBP1, and BASP1 protein expression”. In: *Molecular Psychiatry* 14.6 (2009), pp. 601–613. DOI: [10.1038/mp.2008.7](https://doi.org/10.1038/mp.2008.7).
- [14] A. Vincentelli. “Platform-based design”. In: (2002).
- [15] A. Yuan, E. Garcia, H. Zhu, and S. Samtani. “Depressive behavior detection using sensor signal data: An attention-based privacy-preserving approach”. In: 2025. DOI: [10.24251/HICSS.2025.049](https://doi.org/10.24251/HICSS.2025.049).
- [16] Y. Abe. “Fatigue as a core symptom of insomnia”. In: *Innov. Clin. Neurosci.* 9.2 (2012), pp. 10–11.
- [17] Agency for Healthcare Research and Quality (US). *2022 National Healthcare Quality and Disparities Report*. 2022.
- [18] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5-TR*. Fifth Edition. 2022.
- [19] B. Ganji, A. Rezaee, S. Adabi, and A. Movaghar. “Model verification of real-time and distributed stream processing architecture”. In: *Computing* 107.17 (2025). DOI: [10.1007/s00607-024-01384-w](https://doi.org/10.1007/s00607-024-01384-w).
- [20] B. Löwe, R. L. Spitzer, K. Gräfe, K. Kroenke, A. Quenter, S. Zipfel, C. Buchholz, S. Witte, and W. Herzog. “Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians’ diagnoses”. In: *Journal of Affective Disorders* 78.2 (2004), pp. 131–140. DOI: [10.1016/S0165-0327\(02\)00237-9](https://doi.org/10.1016/S0165-0327(02)00237-9).
- [21] B. McFee, M. McVicar, D. Faronbi, I. Roman, M. Gover, S. Balke, S. Seyfarth, A. Malek, C. Raffel, V. Lostanlen, B. van Niekirk, D. Lee, F. Cwitkowitz, F. Zalkow, O. Nieto, D. Ellis, J. Mason, K. Lee, B. Steers, D. Südholt, et al. *librosa/librosa: 0.11.0 (0.11.0)*. 2025. DOI: [10.5281/zenodo.15006942](https://doi.org/10.5281/zenodo.15006942). URL: <https://doi.org/10.5281/zenodo.15006942>.
- [22] B. Mishra and A. Kertesz. “The use of MQTT in M2M and IoT systems: A survey”. In: *IEEE Access* 8 (2020), pp. 201071–201086. DOI: [10.1109/ACCESS.2020.3035849](https://doi.org/10.1109/ACCESS.2020.3035849).
- [23] B. O’Dea, T. A. Braund, P. J. Batterham, M. E. Larsen, N. Glozier, and A. E. Whitton. “Reading between the lines: Identifying the linguistic markers of anhedonia for the stratification of depression”. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 2024. ISBN: 9798400703300. DOI: [10.1145/3613904.3642478](https://doi.org/10.1145/3613904.3642478).

- [24] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini. “The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, sincerity & native language”. In: *Proc. Interspeech 2016*. 2016, pp. 2001–2005. DOI: [10.21437/Interspeech.2016-129](https://doi.org/10.21437/Interspeech.2016-129).
- [25] B. Stasak, J. Epps, N. Cummins, and R. Goecke. “An investigation of emotional speech in depression classification”. In: *Proceedings of Interspeech 2016*. 2016, pp. 485–489.
- [26] B. T. Atmaja, A. Sasou, and M. Akagi. “Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion”. In: *Speech Communication* 140 (2022). DOI: [10.1016/j.specom.2022.03.002](https://doi.org/10.1016/j.specom.2022.03.002).
- [27] B. W. Schuller and A. M. Batliner. “Front matter”. In: *Computational Paralinguistics*. 2013, pp. i–xxi. ISBN: 9781118706664. DOI: <https://doi.org/10.1002/9781118706664.fmatter>.
- [28] M. B. Bedir. *Understanding exponential moving averages (EMA) and their applications in Python*. 2024.
- [29] D. Bolton. “Overdiagnosis problems in the DSM-IV and the new DSM-5: Can they be resolved by the distress—impairment criterion?” In: *The Canadian Journal of Psychiatry* 58.11 (2013), pp. 612–617. DOI: [10.1177/070674371305801106](https://doi.org/10.1177/070674371305801106).
- [30] Google Brain. *Inception V3 Model*. 2021.
- [31] C. A. Stewart, D. G. V. Mitchell, P. A. MacDonald, S. G. Pasternak, P. F. Tremblay, and E. C. Finger. “The nonverbal expression of guilt in healthy adults”. In: *Scientific Reports* 14.1 (2024), p. 10607. DOI: [10.1038/s41598-024-60980-0](https://doi.org/10.1038/s41598-024-60980-0).
- [32] C. A. Woody, A. J. Ferrari, D. J. Siskind, H. A. Whiteford, M. G. Harris. “A systematic review and meta-regression of the prevalence and incidence of perinatal depression”. In: *Journal of Affective Disorders* 219 (2017), pp. 86–92. DOI: [10.1016/j.jad.2017.05.003](https://doi.org/10.1016/j.jad.2017.05.003).
- [33] C. Lin, P. Hu, H. Su, S. Li, J. Mei, J. Zhou, H. Leung. “SenseMood: Depression detection on social media”. In: *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 2020, pp. 407–411. ISBN: 9781450370875. DOI: [10.1145/3372278.3391932](https://doi.org/10.1145/3372278.3391932).
- [34] C. Montag, J. D. Elhai, and P. Dagum. “On blurry boundaries when defining digital biomarkers: How much biology needs to be in a digital biomarker?” In: *Frontiers in Psychiatry* 12 (2021), p. 740292. DOI: [10.3389/fpsy.2021.740292](https://doi.org/10.3389/fpsy.2021.740292).
- [35] C. Richardson. *Microservices patterns*. 2019. ISBN: 9781617294549.
- [36] C. Shin, S. H. Lee, K. M. Han, H. K. Yoon, and C. Han. “Comparison of the usefulness of the PHQ-8 and PHQ-9 for screening for major depressive disorder: Analysis of psychiatric outpatient data”. In: *Psychiatry Investigation* 16.4 (2019), pp. 300–305. DOI: [10.30773/pi.2019.02.01](https://doi.org/10.30773/pi.2019.02.01).
- [37] Centers for Disease Control and Prevention. *Child and teen BMI calculator*. Accessed: 27-04-2025. 2024. URL: <https://www.cdc.gov/bmi/child-teen-calculator/index.html>.
- [38] S.P. Chand and H. Arif. *Depression*. 2023.
- [39] M. Cohn. *User stories applied: For agile software development*. 2004.

- [40] SciPy Community. *scipy.stats.pearsonr*. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html>. Accessed: 23-06-2025. 2024.
- [41] USC Institute for Creative Technologies. *DAIC-WOZ database & extended DAIC database*. <https://dcapswoz.ict.usc.edu/>. Accessed: 10-06-2025. June 2025.
- [42] D. A. Regier, W. E. Narrow, D. E. Clarke, H. C. Kraemer, S. J. Kuramoto, E. A. Kuhl, D. J. Kupfer. “DSM-5 field trials in the United States and Canada, part II: Test-retest reliability of selected categorical diagnoses”. In: *American Journal of Psychiatry* 170.1 (2013), pp. 59–70. DOI: [10.1176/appi.ajp.2012.12070999](https://doi.org/10.1176/appi.ajp.2012.12070999).
- [43] D. Ameller, C. Ayala, J. Cabot, X. Franch. “Non-functional requirements in architectural decision-making”. In: *IEEE Software* 30 (2012), pp. 61–67. DOI: [10.1109/MS.2012.176](https://doi.org/10.1109/MS.2012.176).
- [44] D. Bennabi, P. Vandel, C. Papaxanthis, T. Pozzo, and E. Haffen. “Psychomotor retardation in depression: a systematic review of diagnostic, pathophysiologic, and therapeutic implications”. In: *BioMed Research International* 2013 (2013), p. 158746. DOI: [10.1155/2013/158746](https://doi.org/10.1155/2013/158746).
- [45] D. DeVault, et al. “SimSensei kiosk: A virtual human interviewer for healthcare decision support”. In: *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems (AAMAS '14)*. 2014, pp. 1061–1068. ISBN: 9781450327381.
- [46] D. F. Santomauro, A. M. Mantilla Herrera, J. Shadid, et al. “Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic”. In: *The Lancet* 398.10312 (2021), pp. 1700–1712. ISSN: 0140-6736. DOI: [https://doi.org/10.1016/S0140-6736\(21\)02143-7](https://doi.org/10.1016/S0140-6736(21)02143-7).
- [47] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and D. 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](https://doi.org/10.1109/10.846676).
- [48] D. João, X. Wang, and E. Guerra. “How to design future-ready microservices? Analyzing microservice patterns for adaptability”. In: *Proceedings of the 28th European Conference on Pattern Languages of Programs*. 2024. ISBN: 9798400700408. DOI: [10.1145/3628034.3628046](https://doi.org/10.1145/3628034.3628046).
- [49] D. K. Saha, T. Hossain, M. Safran, et al. “Ensemble of hybrid model based technique for early detecting of depression based on SVM and neural networks”. In: *Scientific Reports* 14 (2024), p. 25470. DOI: [10.1038/s41598-024-77193-0](https://doi.org/10.1038/s41598-024-77193-0).
- [50] D. M. Low, K. H. Bentley, S. S. Ghosh. “Automated assessment of psychiatric disorders using speech: A systematic review”. In: *Laryngoscope Investigative Otolaryngology* 5.1 (2020), pp. 96–116. DOI: <https://doi.org/10.1002/lio2.354>.
- [51] D. Martins-de-Souza, E. Dias-Neto, A. Schmitt, P. Falkai, P. Gormanns, G. Mac-carrone, et al. “Proteome analysis of schizophrenia brain tissue”. In: *World Journal of Biological Psychiatry* 11.2 (2010), pp. 110–120. DOI: [10.3109/15622970903490626](https://doi.org/10.3109/15622970903490626).
- [52] C. Dilmegani. *TinyML (EdgeAI) in 2025: Machine learning at the edge*. Accessed: 23-05-2025. 2025. URL: <https://research.aimultiple.com/tinyml/>.

- [53] E. Dafna, A. Tarasiuk, and Y. Zigel. “Sleep staging using nocturnal sound analysis”. In: *Scientific Reports* 8.1 (2018), p. 13474. DOI: [10.1038/s41598-018-31748-0](https://doi.org/10.1038/s41598-018-31748-0).
- [54] E. I. Kvig and S. Nilssen. “Does method matter? Assessing the validity and clinical utility of structured diagnostic interviews among a clinical sample of first-admitted patients with psychosis: A replication study”. In: *Frontiers in Psychiatry* Volume 14 - 2023 (2023). ISSN: 1664-0640. DOI: [10.3389/fpsy.2023.1076299](https://doi.org/10.3389/fpsy.2023.1076299).
- [55] E. Mahon and M. E. Lachman. “Voice biomarkers as indicators of cognitive changes in middle and later adulthood”. In: *Neurobiology of Aging* 119 (2022), pp. 22–35. DOI: [10.1016/j.neurobiolaging.2022.06.010](https://doi.org/10.1016/j.neurobiolaging.2022.06.010).
- [56] E. R. Simon-Thomas, D. J. Keltner, D. Sauter, L. Sinicropi-Yao, and A. Abramson. “The voice conveys specific emotions: Evidence from vocal burst displays”. In: *Emotion* 9.6 (2009), pp. 838–846. DOI: [10.1037/a0017810](https://doi.org/10.1037/a0017810).
- [57] E. Szabadi, C. M. Bradshaw, and J. A. O. Besson. “Elongation of pause-time in speech: A simple, objective measure of motor retardation in depression”. In: *British Journal of Psychiatry* 129.6 (1976), pp. 592–597. DOI: [10.1192/bjp.129.6.592](https://doi.org/10.1192/bjp.129.6.592).
- [58] E. Thoret, T. Andrillon, C. Gauriau, D. Léger, and D. Pressnitzer. “Sleep deprivation detected by voice analysis”. In: *PLOS Computational Biology* 20.2 (2024), pp. 1–24. DOI: [10.1371/journal.pcbi.1011849](https://doi.org/10.1371/journal.pcbi.1011849).
- [59] B. B. Eye. “Depression analysis”. In: *Kaggle* (2020).
- [60] F. Cacheda, D. Fernandez, F. J. Novoa, and V. Carneiro. “Early detection of depression: Social network analysis and random forest techniques”. In: *Journal of Medical Internet Research* 21.6 (2019), e12554. DOI: [10.2196/12554](https://doi.org/10.2196/12554).
- [61] F. Eyben, F. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. 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](https://doi.org/10.1109/TAFFC.2015.2457417).
- [62] F. Eyben, M. Wöllmer, and B. Schuller. “Opensmile: the munich versatile and fast open-source audio feature extractor”. In: 2010, pp. 1459–1462. ISBN: 9781605589336. DOI: [10.1145/1873951.1874246](https://doi.org/10.1145/1873951.1874246).
- [63] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski. “Automatic modelling of depressed speech: Relevant features and relevance of gender”. In: *Proceedings of Interspeech*. 2014, pp. 1248–1252.
- [64] F. Jungbauer, R. Hülse, F. Lu, S. Ludwig, V. Held, N. Rotter, and A. Schell. “Case report: Bilateral palsy of the vocal cords after COVID-19 infection”. In: *Frontiers in Neurology* Volume 12 (2021). ISSN: 1664-2295. DOI: [10.3389/fneur.2021.619545](https://doi.org/10.3389/fneur.2021.619545).
- [65] F. L. Dias-Audibert, L. C. Navarro, D. N. de Oliveira, J. Delafiori, C. F. O. R. Melo, T. M. Guerreiro, F. T. Rosa, D. L. Petenuci, M. A. E. Watanabe, L. A. Velloso, A. R. Rocha, R. R. Catharino. “Combining machine learning and metabolomics to identify weight gain biomarkers”. In: *Frontiers in Bioengineering and Biotechnology* 8 (2020), p. 6. DOI: [10.3389/fbioe.2020.00006](https://doi.org/10.3389/fbioe.2020.00006).

- [66] F. Wang, R. Kaushal, and D. Khullar. “Should health care demand interpretable artificial intelligence or accept “black box” medicine?” In: *Annals of Internal Medicine* 172.1 (2020), pp. 59–60. DOI: [10.7326/M19-2548](https://doi.org/10.7326/M19-2548).
- [67] N. El Fadel. “Facial recognition algorithms: A systematic literature review”. In: *J. Imaging* 11.2 (2025), p. 58. DOI: [10.3390/jimaging11020058](https://doi.org/10.3390/jimaging11020058).
- [68] J. Faircloth. “Chapter 7 - Testing enterprise applications”. In: *Penetration Tester’s Open Source Toolkit (Fourth Edition)*. Fourth Edition. 2017, pp. 243–271. ISBN: 978-0-12-802149-1. DOI: <https://doi.org/10.1016/B978-0-12-802149-1.00007-5>.
- [69] G. D. Ruxton and M. Neuhäuser. “When should we use one-tailed hypothesis testing?” In: *Methods in Ecology and Evolution* 1.2 (2010), pp. 114–117. DOI: [10.1111/j.2041-210X.2010.00014.x](https://doi.org/10.1111/j.2041-210X.2010.00014.x).
- [70] G. M. Harari, S. R. Müller, M. S. H. Aung, and P. J. Rentfrow. “Smartphone sensing methods for studying behavior in everyday life”. In: *Current opinion in behavioral sciences* 18 (2017), pp. 83–90. ISSN: 2352-1546. DOI: <https://doi.org/10.1016/j.cobeha.2017.07.018>.
- [71] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu. “Depression detection via harvesting social media: A multimodal dictionary learning solution”. In: *Proceedings of the twenty-sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 2017, pp. 3838–3844. DOI: [10.24963/ijcai.2017/536](https://doi.org/10.24963/ijcai.2017/536).
- [72] G. W. Arana, R. J. Baldessarini, and M. Ornstein. “The dexamethasone suppression test for diagnosis and prognosis in psychiatry”. In: *Archives of General Psychiatry* 42.12 (1985), pp. 1193–1204. DOI: <https://doi.org/10.1001/archpsyc.1985.01790350067012>.
- [73] H. Cervantes and R. Kazman. *Designing software architectures: A practical approach*. 2016. ISBN: 978-0134390789.
- [74] H. Foidl, V. Golendukhina, R. Ramler, and M. Felderer. “Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers”. In: *Journal of Systems and Software* 207 (2024), p. 111855. DOI: [10.1016/j.jss.2023.111855](https://doi.org/10.1016/j.jss.2023.111855).
- [75] H. Konno, J. Toyama, M. Shimbo, and K. Murata. “The effect of formant frequency and spectral tilt of unvoiced vowels on their perceived pitch and phonemic quality”. In: *IEICE technical report. Speech* 95.565 (1996), pp. 39–45.
- [76] H. Le-Niculescu, S. M. Kurian, N. Yehyawji, C. Dike, S. D. Patel, H. J. Edenberg, M. T. Tsuang, D. R. Salomon, J. I. Nurnberger, and A. B. Niculescu. “Identifying blood biomarkers for mood disorders using convergent functional genomics”. In: *Molecular Psychiatry* 14.2 (2009), pp. 156–174. DOI: [10.1038/mp.2008.11](https://doi.org/10.1038/mp.2008.11).
- [77] H. Tufail, S. M. Cheema, M. Ali, I. M. Pires, and N. M. Garcia. “Depression detection with convolutional neural networks: A step towards improved mental health care”. In: *Procedia Computer Science* 224 (2023), pp. 544–549. ISSN: 1877-0509. DOI: <https://doi.org/10.1016/j.procs.2023.09.079>.
- [78] H. Vural, M. Koyuncu, and S. Guney. “A systematic literature review on microservices”. In: 2017, pp. 203–217. ISBN: 978-3-319-62406-8. DOI: [10.1007/978-3-319-62407-5_14](https://doi.org/10.1007/978-3-319-62407-5_14).

- [79] H. Wang, M. S. U. Sourav, M. Yang, and J. Zhang. “Classifying mental disorders through clinicians’ subjective approach based on three-way decisions”. In: *Frontiers in Psychology* Volume 14 (2023). ISSN: 1664-1078. DOI: [10.3389/fpsyg.2023.1144826](https://doi.org/10.3389/fpsyg.2023.1144826).
- [80] H. Yu, L. Koban, M. J. Crockett, X. Zhou, and T. D. Wager. “Toward a brain-based bio-marker of guilt”. In: *Neuroscience Insights* 15 (2020), p. 2633105520957638. DOI: [10.1177/2633105520957638](https://doi.org/10.1177/2633105520957638).
- [81] B. H. Hidaka. “Depression as a disease of modernity: explanations for increasing prevalence”. In: *Journal of Affective Disorders* 140.3 (2012), pp. 205–214. DOI: [10.1016/j.jad.2011.12.036](https://doi.org/10.1016/j.jad.2011.12.036).
- [82] I. F. Eiko and R. M. Nesse. “Depression sum-scores don’t add up: why analyzing specific depression symptoms is essential”. In: *BMC Medicine* 13.1 (2015), p. 72. DOI: [10.1186/s12916-015-0325-4](https://doi.org/10.1186/s12916-015-0325-4).
- [83] I. Gómez-Gómez, I. Benítez, J. Bellón, et al. “Utility of PHQ-2, PHQ-8 and PHQ-9 for detecting major depression in primary health care: a validation study in Spain”. In: *Psychological Medicine* 53.12 (2023), pp. 5625–5635. DOI: [10.1017/S0033291722002835](https://doi.org/10.1017/S0033291722002835).
- [84] I. M. Putrama and P. Martinek. “Heterogeneous data integration: Challenges and opportunities”. In: *Data in Brief* 56 (2024), p. 110853. ISSN: 2352-3409. DOI: <https://doi.org/10.1016/j.dib.2024.110853>.
- [85] I. Sommerville, and Pearson Studium. *Modernes software-engineering: Entwurf und Entwicklung von Softwareprodukten*. 2020. ISBN: 9783868943962.
- [86] C. Iber. *The AASM manual for the scoring of sleep and associated events: Rules, terminology, and technical specifications*. 1st. 2007.
- [87] IBM. *What is a REST API?* <https://www.ibm.com/think/topics/rest-apis>. Accessed: 23-06-2025. 2025.
- [88] Streamlit Inc. <https://streamlit.io/>. Accessed: 01-06-2025. 2025.
- [89] R. Iyer and D. Meyer. “Detection of suicide risk using vocal characteristics: Systematic review”. In: *JMIR Biomed Eng* 7.2 (2022), e42386. ISSN: 2561-3278. DOI: [10.2196/42386](https://doi.org/10.2196/42386).
- [90] J. A. Hall, J. A. Harrigan, and R. Rosenthal. “Nonverbal behavior in clinician—patient interaction”. In: *Applied and Preventive Psychology* 4.1 (1995), pp. 21–37. ISSN: 0962-1849. DOI: [https://doi.org/10.1016/S0962-1849\(05\)80049-6](https://doi.org/10.1016/S0962-1849(05)80049-6).
- [91] J. Acurio, C. Celis, J. Perez and C. Escudero. “Acoustic parameters and salivary IL-6 levels in overweight and obese teachers”. In: *Journal of Voice* 28.5 (2014), pp. 574–581. DOI: [10.1016/j.jvoice.2014.03.002](https://doi.org/10.1016/j.jvoice.2014.03.002).
- [92] J. C. Ballenger, J. R. Davidson, Y. Lecrubier, D. J. Nutt, D. Goldberg, K. M. Magruder, H. C. Schulberg, A. Tylee, and H. U. Wittchen. “Consensus statement on the primary care management of depression from the International Consensus Group on Depression and Anxiety”. In: *The Journal of Clinical Psychiatry* 60.Suppl 7 (1999), pp. 54–61.

- [93] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. “Detecting depression from facial actions and vocal prosody”. In: *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. 2009, pp. 1–7. DOI: [10.1109/ACII.2009.5349358](https://doi.org/10.1109/ACII.2009.5349358).
- [94] J. Feightner and G. Worrall. “Early detection of depression”. In: *Preventing disease: Beyond the rhetoric*. 1990, pp. 118–128. ISBN: 978-1-4612-3280-3. DOI: [10.1007/978-1-4612-3280-3_15](https://doi.org/10.1007/978-1-4612-3280-3_15).
- [95] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency. “The distress analysis interview corpus of human and computer interviews”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. 2014, pp. 3123–3128.
- [96] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami. “Internet of Things (IoT): A vision, architectural elements, and future directions”. In: *Future Generation Computer Systems* 29.7 (2013), pp. 1645–1660. ISSN: 0167-739X. DOI: [10.1016/j.future.2013.01.010](https://doi.org/10.1016/j.future.2013.01.010).
- [97] J. Huang, P. Guo, S. Zhang, M. Ji, and R. An. “Use of deep neural networks to predict obesity with short audio recordings: Development and usability study”. In: *JMIR AI* 3 (2024), e54885. DOI: [10.2196/54885](https://doi.org/10.2196/54885).
- [98] J. J. G. Meilán, F. Martínez-Sánchez, I. Martínez-Nicolás, T. E. Llorente, and J. Carro. “Changes in the rhythm of speech difference between people with nondegenerative mild cognitive impairment and with preclinical dementia”. In: *Behavioural Neurology* 2020.1 (2020), p. 4683573. DOI: <https://doi.org/10.1155/2020/4683573>.
- [99] J. Pushpa and S. A. Kalyani. “Chapter three - Using fog computing/edge computing to leverage digital twin”. In: *Advances in Computers*. Vol. 117. 2020, pp. 51–77. ISBN: 9780128187562. DOI: [10.1016/bs.adcom.2019.09.003](https://doi.org/10.1016/bs.adcom.2019.09.003).
- [100] J. Shreffler, M.-R. Huecker. “Type I and Type II errors and statistical power”. In: *StatPearls [Internet]*. 2025.
- [101] J. T. Schousboe, T. N. Vo, A. M. Kats, L. Langsetmo, S. J. Diem, B. C. Taylor, E. S. Strotmeyer, and K. E. Ensrud. “Depressive Symptoms and Total Healthcare Costs: Roles of Functional Limitations and Multimorbidity”. In: *Journal of the American Geriatrics Society* 67.8 (2019), pp. 1596–1603. DOI: [10.1111/jgs.15881](https://doi.org/10.1111/jgs.15881).
- [102] J. Vaissière. “Perception of intonation”. In: *The Handbook of Speech Perception*. 2005, pp. 236–263.
- [103] J. Venable, J. Pries-Heje, and R. Baskerville. “A comprehensive framework for evaluation in design science research”. In: *LNCS* 7286 (2012), pp. 423–438. DOI: [10.1007/978-3-642-29863-9_31](https://doi.org/10.1007/978-3-642-29863-9_31).
- [104] J. Williamson, D. Young, A. Nierenberg, J. Niemi, B. Helfer, and T. Quatieri. “Tracking depression severity from audio and video based on speech articulatory coordination”. In: *Computer Speech & Language* 55 (2019). DOI: [10.1016/j.csl.2018.08.004](https://doi.org/10.1016/j.csl.2018.08.004).

- [105] K. A. L. Govindasamy and N. Palanichamy. “Depression detection using machine learning techniques on Twitter data”. In: *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. 2021, pp. 960–966. DOI: [10.1109/ICICCS51141.2021.9432203](https://doi.org/10.1109/ICICCS51141.2021.9432203).
- [106] K. Fodor, Z. Balogh, and G. Molnár. “Real-time emotion recognition in smart homes”. In: 2023. DOI: [10.1109/SACI58269.2023.10158664](https://doi.org/10.1109/SACI58269.2023.10158664).
- [107] K. J. Kallail and F. W. Emanuel. “An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects”. In: *Journal of Phonetics* 12.2 (1984), pp. 175–186. ISSN: 0095-4470. DOI: [https://doi.org/10.1016/S0095-4470\(19\)30864-2](https://doi.org/10.1016/S0095-4470(19)30864-2).
- [108] K. Kaczmarek-Majer, M. Dominiak, A. Z. Antosik, O. Hryniewicz, O. Kamińska, K. Opara, J. Owsinski, W. Radziszewska, M. Sochacka, and L. Święcicki. “Acoustic features from speech as markers of depressive and manic symptoms in bipolar disorder: A prospective study”. In: *Acta Psychiatrica Scandinavica* 151.3 (2025), pp. 358–374. DOI: <https://doi.org/10.1111/acps.13735>.
- [109] K. Kroenke, R. L. Spitzer, and J. B. W. Williams. “The PHQ-9: validity of a brief depression severity measure”. In: *Journal of General Internal Medicine* 16.9 (2001), pp. 606–613. DOI: [10.1046/j.1525-1497.2001.016009606.x](https://doi.org/10.1046/j.1525-1497.2001.016009606.x).
- [110] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad. “The PHQ-8 as a measure of current depression in the general population”. In: *Journal of Affective Disorders* 114.1-3 (2009), pp. 163–173. DOI: [10.1016/j.jad.2008.06.026](https://doi.org/10.1016/j.jad.2008.06.026).
- [111] K. M. Smith, P. F. Renshaw, and J. Bilello. “The diagnosis of depression: current and emerging methods”. In: *Comprehensive Psychiatry* 54.1 (2013), pp. 1–6. DOI: [10.1016/j.comppsy.2012.06.006](https://doi.org/10.1016/j.comppsy.2012.06.006).
- [112] K. Mao, Y. Wu, and J. Chen. “A systematic review on automated clinical depression diagnosis”. In: *npj Mental Health Research* 2 (2023), p. 20. DOI: [10.1038/s44184-023-00040-z](https://doi.org/10.1038/s44184-023-00040-z).
- [113] K. Min, J. Yoon, M. Kang, D. Lee, E. Park, and J. Han. “Detecting depression on video logs using audiovisual features”. In: *Humanities and Social Sciences Communications* 10.1 (2023), p. 788. DOI: [10.1057/s41599-023-02313-6](https://doi.org/10.1057/s41599-023-02313-6).
- [114] K. S. Jacob and V. Patel. “Classification of mental disorders: a global mental health perspective”. In: *The Lancet* 383.9926 (2014), pp. 1433–1435. DOI: [10.1016/S0140-6736\(13\)62382-X](https://doi.org/10.1016/S0140-6736(13)62382-X).
- [115] K. Smith, P. Renshaw, and J. Bilello. “The diagnosis of depression: Current and emerging methods”. In: *Comprehensive psychiatry* 54.1 (2013), pp. 1–6. DOI: [10.1016/j.comppsy.2012.06.006](https://doi.org/10.1016/j.comppsy.2012.06.006).
- [116] K. Tjaden and G. E. Wilding. “Rate and loudness manipulations in dysarthria”. In: *Journal of Speech, Language, and Hearing Research* 47.4 (2004), pp. 766–783. DOI: [10.1044/1092-4388\(2004/058\)](https://doi.org/10.1044/1092-4388(2004/058)).
- [117] K. Tjaden, G. Wilding. “Speech and pause characteristics associated with voluntary rate reduction in Parkinson’s disease and multiple sclerosis”. In: *Journal of Communication Disorders* 44.6 (2011), pp. 655–665. DOI: [10.1016/j.jcomdis.2011.06.003](https://doi.org/10.1016/j.jcomdis.2011.06.003).

- [118] D. Keltner. “Evidence for the distinctness of embarrassment, shame, and guilt: A study of recalled antecedents and facial expressions of emotion”. In: *Cognition and Emotion* 10.2 (1996), pp. 155–172. DOI: [10.1080/026999396380312](https://doi.org/10.1080/026999396380312).
- [119] M. Kleppmann. *Designing data-intensive applications*. First. 2017. ISBN: 978-1-449-37332-0.
- [120] E. Kraepelin. “Manic depressive insanity and paranoia”. In: *The Journal of Nervous and Mental Disease* 53.4 (1921), p. 350.
- [121] L. A. Paige, M. W. Mitchell, K. R. R. Krishnan, R. Kaddurah-Daouk, and D. C. Steffens. “A preliminary metabolomic analysis of older adults with and without depression”. In: *International Journal of Geriatric Psychiatry* 22.5 (2007), pp. 418–423. DOI: [10.1002/gps.1690](https://doi.org/10.1002/gps.1690).
- [122] L. Cui, S. Li, S. Wang, X. Wu, Y. Liu, W. Yu, Y. Wang, Y. Tang, M. Xia, and B. Li. “Major depressive disorder: hypothesis, mechanism, prevention and treatment”. In: *Signal Transduction and Targeted Therapy* 9.30 (2024).
- [123] L. Kalevo, M. Huotari, R. Kiviniemi, A. Bruna, S. Myllymaa, T. Himanen, and L. Nikkonen. “Self-applied electrode set provides a clinically feasible solution enabling EEG recording in home sleep apnea testing”. In: *IEEE Access* 10 (2022), pp. 60633–60642. DOI: [10.1109/ACCESS.2022.3178189](https://doi.org/10.1109/ACCESS.2022.3178189).
- [124] L. Rachakonda, A. Kothari, S. Mohanty, E. Kougianos, and M. Ganapathiraju. “Stress-log: An IoT-based smart system to monitor stress-eating”. In: 2019, pp. 1–6. DOI: [10.1109/ICCE.2019.8661959](https://doi.org/10.1109/ICCE.2019.8661959).
- [125] L. S. A. Low, N. C. Maddage, M. Lech, and N. Allen. “Mel frequency cepstral feature and Gaussian Mixtures for modeling clinical depression in adolescents”. In: *2009 8th IEEE International Conference on Cognitive Informatics*. 2009, pp. 346–350. DOI: [10.1109/COGINF.2009.5250714](https://doi.org/10.1109/COGINF.2009.5250714).
- [126] L.-V. Herm, K. Heinrich, J. Wanner, and C. Janiesch. “Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability”. In: *International Journal of Information Management* 69 (2023), p. 102538. ISSN: 0268-4012. DOI: <https://doi.org/10.1016/j.ijinfomgt.2022.102538>.
- [127] APA Task Force on Laboratory Tests in Psychiatry. “The dexamethasone suppression test: an overview of its current status in psychiatry”. In: *American Journal of Psychiatry* 144.10 (1987), pp. 1253–1262.
- [128] Jonas Länzlinger. *Depression Detection GitHub Repository*. <https://github.com/jonaslanzlinger/depression-detection>. Accessed: 21-06-2025. 2025.
- [129] A. T. Latinjak. “The underlying structure of emotions: A tri-dimensional model of core affect and emotion concepts for sports”. In: *Revista Iberoamericana de Psicología del Ejercicio y el Deporte* 7.1 (2012), pp. 71–88.
- [130] J. Laver. *Principles of phonetics*. 1994. ISBN: 9780521452129.
- [131] D. Linthicum. “Megatrend alert: The rise of ubiquitous computing”. In: *InfoWorld* (2023).
- [132] F. W. Lohoff. “Overview of the genetics of major depressive disorder”. In: *Current Psychiatry Reports* 12.6 (2010), pp. 539–546. DOI: [10.1007/s11920-010-0150-6](https://doi.org/10.1007/s11920-010-0150-6).

- [133] Lumen Learning. *Clinical assessments and the mental status examination*. Accessed: 05-06-2025. 2025. URL: <https://courses.lumenlearning.com/wm-abnormalpsych/chapter/clinical-assessments-and-the-mental-status-examination/>.
- [134] M. A. Imran, Z. Ahmed, L. Zhang, Q. H. Abbasi. “Grand challenges in IoT and sensor networks”. In: *Frontiers in Communications and Networks* 1 (2020). ISSN: 2673-530X. DOI: [10.3389/frcmn.2020.619452](https://doi.org/10.3389/frcmn.2020.619452).
- [135] M. A. Marty and D. L. Segal. “DSM-5”. In: *The encyclopedia of clinical psychology*. 2015, pp. 1–6. ISBN: 9781118625392. DOI: <https://doi.org/10.1002/9781118625392.wbecp308>.
- [136] M. B. First, W. Gaebel, M. Maj, D. J. Stein, C. S. Kogan, J. B. Saunders, V. B. Poznyak, O. Gureje, R. Lewis-Fernández, A. Maercker, C. R. Brewin, M. Cloitre, A. Claudino, K. M. Pike, G. Baird, D. Skuse, R. B. Krueger, P. Briken, J. D. Burke, J. E. Lochman, G. M. Reed, et al. “An organization- and category-level comparison of diagnostic requirements for mental disorders in ICD-11 and DSM-5”. In: *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)* 20.1 (2021), pp. 34–51. DOI: [10.1002/wps.20825](https://doi.org/10.1002/wps.20825).
- [137] M. Broadbent, M. Medina Grespan, K. Axford, X. Zhang, V. Srikumar, B. Kious, and Z. Imel. “A machine learning approach to identifying suicide risk among text-based crisis counseling encounters”. In: *Frontiers in Psychiatry* 14 (2023), p. 1110527. DOI: [10.3389/fpsy.2023.1110527](https://doi.org/10.3389/fpsy.2023.1110527).
- [138] M. Daun, A. M. Grubb, V. Stenkova, et al. “A systematic literature review of requirements engineering education”. In: *Requirements Engineering* 28 (2023), pp. 145–175. DOI: [10.1007/s00766-022-00381-9](https://doi.org/10.1007/s00766-022-00381-9).
- [139] M. de Zambotti, N. Cellini, A. Goldstone, I. M. Colrain, and F. C. Baker. “Wearable sleep technology in clinical and research settings”. In: *Medicine and Science in Sports and Exercise* 51.7 (2019), pp. 1538–1557. DOI: [10.1249/MSS.0000000000001947](https://doi.org/10.1249/MSS.0000000000001947).
- [140] M. El Ayadi, M. S. Kamel, and F. Karray. “Survey on speech emotion recognition: Features, classification schemes, and databases”. In: *Pattern Recognition* 44.3 (2011), pp. 572–587. ISSN: 0031-3203. DOI: <https://doi.org/10.1016/j.patcog.2010.09.020>.
- [141] M. Gamer, H.-G. Rill, G. Vossel, and H. Gödert. “Psychophysiological and vocal measures in the detection of guilty knowledge”. In: *International journal of psychophysiology : official journal of the International Organization of Psychophysiology* 60 (2006), pp. 76–87. DOI: [10.1016/j.ijpsycho.2005.05.006](https://doi.org/10.1016/j.ijpsycho.2005.05.006).
- [142] M. Ingle, M. Sharma, K. Kumar, P. Kumar, A. Bhurane, H. Elphick, D. Joshi, U. R. Acharya. “A systematic review on automatic identification of insomnia”. In: *Physiological Measurement* (2024). DOI: [10.1088/1361-6579/ad2059](https://doi.org/10.1088/1361-6579/ad2059).
- [143] M. K. Pichora-Fuller and K. Dupuis. *Toronto emotional speech set (TESS)*. Version V1. 2020. DOI: [10.5683/SP2/E8H2MF](https://doi.org/10.5683/SP2/E8H2MF).
- [144] M. Latinus and M. Taylor. “Discriminating male and female voices: Differentiating pitch and gender”. In: *Brain topography* 25 (2011), pp. 194–204. DOI: [10.1007/s10548-011-0207-9](https://doi.org/10.1007/s10548-011-0207-9).

- [145] M. Marino, Y. Li, M. N. Rueschman, J. W. Winkelman, J. M. Ellenbogen, J. M. Solet, H. Dulin, L. F. Berkman, and O. M. Buxton. “Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography”. In: *Sleep* 36.11 (2013), pp. 1747–1755. DOI: [10.5665/sleep.3142](https://doi.org/10.5665/sleep.3142).
- [146] M. Nibuya, S. Morinobu, and R. S. Duman. “Regulation of BDNF and trkB mRNA in rat brain by chronic electroconvulsive seizure and antidepressant drug treatments”. In: *Journal of Neuroscience* 15.11 (1995), pp. 7539–7547. DOI: [10.1523/JNEUROSCI.15-11-07539.1995](https://doi.org/10.1523/JNEUROSCI.15-11-07539.1995).
- [147] M. Rahul, S. Deena, R. Shylesh, and B. Lanitha. “Detecting and analyzing depression: A comprehensive survey of assessment tools and techniques”. In: *2023 International Conference on Inventive Computation Technologies (ICICT)*. 2023, pp. 749–753. DOI: [10.1109/ICICT57646.2023.10134165](https://doi.org/10.1109/ICICT57646.2023.10134165).
- [148] M. Richards and N. Ford. *Fundamentals of software architecture: A modern engineering approach*. 2025. ISBN: 9781098175474.
- [149] M. Ryo, M. C. Rillig. “Statistically reinforced machine learning for nonlinear patterns and variable interactions”. In: *Ecosphere* 8.11 (2017), e01976. DOI: <https://doi.org/10.1002/ecs2.1976>.
- [150] M. S. García-Gutiérrez, F. Navarrete, F. Sala, A. Gasparian, A. Austrich-Olivares, and J. Manzanares. “Biomarkers in Psychiatry: Concept, Definition, Types and Relevance to the Clinical Reality”. In: *Frontiers in Psychiatry* 11 (2020), p. 432. DOI: [10.3389/fpsyt.2020.00432](https://doi.org/10.3389/fpsyt.2020.00432).
- [151] M. Sadeghi, R. Richer, B. Egger, et al. “Harnessing multimodal approaches for depression detection using large language models and facial expressions”. In: *npj Mental Health Research* 3 (2024), p. 66. DOI: [10.1038/s44184-024-00112-8](https://doi.org/10.1038/s44184-024-00112-8).
- [152] M. Shahrum, I. Isa, N. Azyze, H. Nasir. “Development of an IoT-based sleep pattern monitoring system for sleep disorder detection”. In: *Indonesian Journal of Electrical Engineering and Computer Science* 36 (2024), p. 777. DOI: [10.11591/ijeecs.v36.i2.pp777-784](https://doi.org/10.11591/ijeecs.v36.i2.pp777-784).
- [153] M. Söylemez, B. Tekinerdogan, and A. Kolukısa Tarhan. “Challenges and solution directions of microservice architectures: A systematic literature review”. In: *Applied Sciences* 12.11 (2022). ISSN: 2076-3417. DOI: [10.3390/app12115507](https://doi.org/10.3390/app12115507).
- [154] M. V. Gastel, S. Stuijk, S. Overeem, J. P. V. Dijk, M. M. V. Gilst, G. De Haan. “Camera-based vital signs monitoring during sleep - a proof of concept study”. In: *IEEE Journal of Biomedical and Health Informatics* 25.5 (2021), pp. 1409–1418. DOI: [10.1109/JBHI.2020.3045859](https://doi.org/10.1109/JBHI.2020.3045859).
- [155] M. Yildirim, B. N. Gaynes, P. Keskinocak, B. W. Pence, and J. Swann. “The cost-effectiveness of depression screening for the general adult population”. In: *Journal of Affective Disorders* 303 (2022), pp. 306–314. ISSN: 0165-0327. DOI: <https://doi.org/10.1016/j.jad.2022.02.044>.
- [156] J. H. McDonald. *Multiple comparisons*. <https://stats.libretexts.org/@go/page/1753>. Accessed: 23-06-2025. Jan. 2024.
- [157] S. Molau. “Normalization in the acoustic feature space for improved speech recognition”. Dissertation. Aachen, Technische Hochschule, 2003.
- [158] MongoDB, Inc. *MongoDB: The multi-model database platform – homepage*. <https://www.mongodb.com/>. Accessed: 23-06-2025. 2025.

- [159] N. Bains and S. Abdijadid. “Major Depressive Disorder”. In: *StatPearls [Internet]*. Updated 2023 Apr 10. 2025.
- [160] N. Bains and S. Abdijadid. *Major depressive disorder*. 2023.
- [161] N. C. Jacobson, H. Weingarden, and S. Wilhelm. “Digital biomarkers of mood disorders and symptom change”. In: *NPJ Digital Medicine* 2 (2019), p. 3. DOI: [10.1038/s41746-019-0078-0](https://doi.org/10.1038/s41746-019-0078-0).
- [162] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri. “A review of depression and suicide risk assessment using speech analysis”. In: *Speech Communication* 71 (2015), pp. 10–49. ISSN: 0167-6393. DOI: <https://doi.org/10.1016/j.specom.2015.03.004>.
- [163] N. Drakos and R. Moore. *High-level descriptors*. <https://amatria.in/Thesis/html-thesis/node163.html>. Accessed: 30-04-2025. 2004.
- [164] N. Drakos and R. Moore. *Low-level descriptors*. <https://amatria.in/Thesis/html-thesis/node162.html>. Accessed: 30-04-2025. 2004.
- [165] N. Drakos and R. Moore. *The analysis step*. <https://amatria.in/Thesis/html-thesis/node161.html>. Accessed: 23-06-2025. 2004.
- [166] N. Ford, M. Richards, P. Sadalage, and Z. Dehghani. *Software architecture: The hard parts*. 2021. ISBN: 9781492086864.
- [167] N. H. nik nur wahidah, D. M. Wilkes, and R. Salomon. “Timing patterns of speech as potential indicators of near-term suicidal risk”. In: *International Journal of Multidisciplinary and Current Research* (2015).
- [168] N. H. Nik Nur Wahidah, M. D. Wilkes, R. M. Salomon. “Timing patterns of speech as potential indicators of near-term suicidal risk”. In: *Int. J. of Multidisciplinary and Current research* 3 (2015).
- [169] N. M. Adams-Quackenbush, R. Horselenberg, J. Hubert, A. Vrij, and P. van Koppen. “Interview expectancies: awareness of potential biases influences behaviour in interviewees”. In: *Psychiatry, Psychology and Law* 26.1 (2019), pp. 150–166. DOI: [10.1080/13218719.2018.1485522](https://doi.org/10.1080/13218719.2018.1485522).
- [170] National University Library. *Cohen’s d*. Online statistics resource. Accessed: 16-06-2025. June 2025. URL: <https://resources.nu.edu/statsresources/cohensd>.
- [171] S. Newman. *Building microservices*. First. 2015. ISBN: 978-1-491-95035-7.
- [172] P. Norvig. *How to write a spelling corrector*. Online. Accessed: 26-06-2025. 2007. URL: <http://norvig.com/spell-correct.html>.
- [173] O. Hansson, K. Blennow, H. Zetterberg, and J. Dage. “Blood biomarkers for Alzheimer’s disease in clinical practice and trials”. In: *Nature Aging* 3.5 (2023), pp. 506–519. DOI: [10.1038/s43587-023-00403-3](https://doi.org/10.1038/s43587-023-00403-3).
- [174] O. Remes, J. F. Mendes, and P. Templeton. “Biological, psychological, and social determinants of depression: A review of recent literature”. In: *Brain Sciences* 11.12 (2021), p. 1633. DOI: [10.3390/brainsci11121633](https://doi.org/10.3390/brainsci11121633).
- [175] World Health Organization. *Depressive disorder (depression)*. Accessed: 29-04-2025. Mar. 2023. URL: <https://www.who.int/news-room/fact-sheets/detail/depression>.

- [176] P. B. Nalinde and A. Shinde. “Machine learning framework for detection of psychological disorders at OSN”. In: *International Journal of Innovative Technology and Exploring Engineering* 8 (2019), pp. 3293–3298. DOI: [10.35940/ijitee.I8823.0981119](https://doi.org/10.35940/ijitee.I8823.0981119).
- [177] P. Boersma and D. Weenink. *Praat: Doing phonetics by computer*. Accessed: 29-04-2025. University of Amsterdam. URL: <https://www.fon.hum.uva.nl/praat/>.
- [178] P. Donaghy, E. Ennis, M. Mulvenna, R. R. Bond, N. Kennedy, M. McTear, H. O’Connell, N. Blaylock, R. Brueckner. “A review of studies using machine learning to detect voice biomarkers for depression”. In: *Journal of Technology in Behavioral Science* —.— (2024), pp. 1–15. DOI: [10.1007/s41347-024-00454-2](https://doi.org/10.1007/s41347-024-00454-2).
- [179] P. Ekman, W. V. Friesen, and P. Ellsworth. “What are the relative contributions of facial behavior and contextual information to the judgment of emotion?” In: *Emotion in the human face*. 1982, pp. 111–127.
- [180] P. Fusar-Poli, A. Estradé, G. Stanghellini, C. M. Esposito, R. Rosfort, M. Mancini, P. Norman, J. Cullen, M. Adesina, G. B. Jimenez, C. da Cunha Lewin, E. A. Drah, M. Julien, M. Lamba, E. M. Mutura, B. Prawira, A. Sugianto, J. Teresa, L. A. White, S. Damiani, and M. Maj. “The lived experience of depression: a bottom-up review co-written by experts by experience and academics”. In: *World psychiatry : official journal of the World Psychiatric Association (WPA)* 22.3 (2023), pp. 352–365. DOI: [10.1002/wps.21111](https://doi.org/10.1002/wps.21111).
- [181] P. M. Nadar, M. A. Merrill, K. Austin, et al. “The emergence of psychoanalytical electrochemistry: the translation of MDD biomarker discovery to diagnosis with electrochemical sensing”. In: *Translational Psychiatry* 12 (2022), p. 372. DOI: [10.1038/s41398-022-02138-y](https://doi.org/10.1038/s41398-022-02138-y).
- [182] P. Stewart and J. Newell-Price. “The adrenal cortex”. In: 2016, pp. 489–555. ISBN: 9780323297387. DOI: [10.1016/B978-0-323-29738-7.00015-0](https://doi.org/10.1016/B978-0-323-29738-7.00015-0).
- [183] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. “End-to-end multimodal emotion recognition using deep neural networks”. In: *IEEE Journal of Selected Topics in Signal Processing* 11.8 (2017), pp. 1301–1309. DOI: [10.1109/JSTSP.2017.2764438](https://doi.org/10.1109/JSTSP.2017.2764438).
- [184] D. Parikh and S. Sachdev. “Improving the efficiency of spectral features extraction by structuring the audio files”. In: 2020, pp. 1–5. DOI: [10.1109/HYDCON48903.2020.9242729](https://doi.org/10.1109/HYDCON48903.2020.9242729).
- [185] M. Perry. “The exponentially weighted moving average”. In: 2010. ISBN: 9780470400630. DOI: [10.1002/9780470400531.eorms0314](https://doi.org/10.1002/9780470400531.eorms0314).
- [186] Pfizer Inc. *Patient Health Questionnaire (PHQ-9)*. <https://www.phqscreeners.com>. Copyright © 1999 Pfizer Inc. All rights reserved. Reproduced with permission. PRIME-MD is a trademark of Pfizer Inc. 1999.
- [187] Eclipse Mosquitto MQTT Project. *Mosquitto: An open source MQTT broker and client library*. <https://mosquitto.org/>. Accessed 23-06-2025. 2025.
- [188] Q. Yang, N. Li, Y. Liu, S. Wang, J. Ma, J. Wang, P. Liu, Z. He, G. Wang, and L. Feng. “The association between anhedonia and speech features in depression: A cross-sectional study”. In: *General Hospital Psychiatry* 94 (2025), pp. 192–198. DOI: [10.1016/j.genhosppsych.2025.03.013](https://doi.org/10.1016/j.genhosppsych.2025.03.013).

- [189] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong. “A textual-based featuring approach for depression detection using machine learning classifiers and social media texts”. In: *Computers in Biology and Medicine* 135 (2021), p. 104499. ISSN: 0010-4825. DOI: <https://doi.org/10.1016/j.combiomed.2021.104499>.
- [190] R. F. Baumeister, K. D. Vohs, and D. C. Funder. “Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior?” In: *Perspectives on Psychological Science* 2.4 (2007), pp. 396–403. DOI: [10.1111/j.1745-6916.2007.00051.x](https://doi.org/10.1111/j.1745-6916.2007.00051.x).
- [191] R. F. Lin, T. K. Leung, Y. P. Liu, and K. R. Hu. “Disclosing critical voice features for discriminating between depression and insomnia - a preliminary study for developing a quantitative method”. In: *Healthcare* 10.5 (2022), p. 935. DOI: [10.3390/healthcare10050935](https://doi.org/10.3390/healthcare10050935).
- [192] R. Jahan and M. M. Tripathi. “Chapter two - multimodal depression detection using machine learning”. In: *Artificial intelligence, machine learning, and mental health in pandemics*. 2022, pp. 53–72. ISBN: 978-0-323-91196-2. DOI: <https://doi.org/10.1016/B978-0-323-91196-2.00005-3>.
- [193] R. Kazman and L. Bass. *Toward deriving software architectures from quality attributes*. Tech. rep. CMU/SEI-94-TR-010. 1994.
- [194] R. M. Voss and J. M. Das. “Mental Status Examination”. In: *StatPearls*. 2025.
- [195] R. Nagumo, Y. Zhang, Y. Ogawa, M. Hosokawa, K. Abe, T. Ukeda, S. Sumi, S. Kurita, S. Nakakubo, S. Lee, T. Doi, and H. Shimada. “Automatic detection of cognitive impairments through acoustic analysis of speech”. In: *Current Alzheimer Research* 17.1 (2020), pp. 60–68. DOI: [10.2174/1567205017666200213094513](https://doi.org/10.2174/1567205017666200213094513).
- [196] R. Watson, K. Harvey, C. McCabe, and S. Reynolds. “Understanding anhedonia: a qualitative study exploring loss of interest and pleasure in adolescent depression”. In: *European Child & Adolescent Psychiatry* 29.4 (2020), pp. 489–499. DOI: <https://doi.org/10.1007/s00787-019-01364-y>.
- [197] R. Yadav, Y. Kumar, and R. Patil. “Study of data ingestion tools”. In: *Shodh Sanchar Bulletin* 10 (2023), pp. 246–253.
- [198] *Resemblyzer: A Python package to analyze and compare voices with deep learning*. <https://github.com/resemble-ai/Resemblyzer>. Accessed: 23-05-2025. 2025.
- [199] D. C. Rettew. “Commentary: Diagnoses from clinical evaluations and standardized diagnostic interviews don’t agree”. In: *Psychiatry Times* (2010).
- [200] S. A. Almaghrabi, S. R. Clark, and M. Baumert. “Bio-acoustic features of depression: A review”. In: *Biomedical Signal Processing and Control* 85 (2023), p. 105020. ISSN: 1746-8094. DOI: <https://doi.org/10.1016/j.bspc.2023.105020>.
- [201] S. Adams, K. Reyno-Briscoe, and L. Hutchinson. “Acoustic correlates of monotone speech in Parkinson’s disease”. In: *Canadian Acoustics* 26.3 (1998), pp. 86–87.
- [202] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker. “Detecting depression: A comparison between spontaneous and read speech”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 7547–7551. DOI: [10.1109/ICASSP.2013.6639130](https://doi.org/10.1109/ICASSP.2013.6639130).
- [203] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python*. 2009.

- [204] S. Fara, S. Gorla, E. Molimpakis, and N. Cummins. *Speech and the n-Back task as a lens into depression. How combining both may allow us to isolate different core symptoms of depression*. 2022. arXiv: [2204.00088](https://arxiv.org/abs/2204.00088) [cs.SD]. URL: <https://arxiv.org/abs/2204.00088>.
- [205] S. Graf, T. Herbig, M. Buck, et al. “Features for voice activity detection: a comparative analysis”. In: *EURASIP J. Adv. Signal Process* 91 (2015). DOI: [10.1186/s13634-015-0277-z](https://doi.org/10.1186/s13634-015-0277-z).
- [206] S. H. M. Peersmann, M. A. Grootenhuys, A. van Straten, W. J. E. Tissing, F. Abbink, A. C. H. de Vries, J. Loonen, H. J. H. van der Pal, G. J. L. Kaspers, and R. R. L. van Litsenburg. “Insomnia symptoms and daytime fatigue co-occurrence in adolescent and young adult childhood cancer patients in follow-up after treatment: Prevalence and associated risk factors”. In: *Cancers* 14.14 (2022), p. 3316. DOI: [10.3390/cancers14143316](https://doi.org/10.3390/cancers14143316).
- [207] S. Hassan, R. Bahsoon, and R. Buyya. “Systematic scalability analysis for microservices granularity adaptation design decisions”. In: *Software: Practice and Experience* 52.6 (2022), pp. 1378–1401. DOI: <https://doi.org/10.1002/spe.3069>.
- [208] S. Homan, M. Gabi, N. Klee, S. Bachmann, A.-M. Moser, M. Duri’, S. Michel, A.-M. Bertram, A. Maatz, G. Seiler, E. Stark, and B. Kleim. “Linguistic features of suicidal thoughts and behaviors: A systematic review”. In: *Clinical Psychology Review* 95 (2022), p. 102161. ISSN: 0272-7358. DOI: <https://doi.org/10.1016/j.cpr.2022.102161>.
- [209] S. J. Kim, S. Kim, S. Jeon, E. B. Leary, F. Barwick, and E. Mignot. “Factors associated with fatigue in patients with insomnia”. In: *Journal of Psychiatric Research* 117 (2019), pp. 24–30. DOI: [10.1016/j.jpsychires.2019.06.021](https://doi.org/10.1016/j.jpsychires.2019.06.021).
- [210] S. L. Redburn. “Weight fluctuation and the voice: A survey of effects from the singer’s perspective”. Electronic Theses and Dissertations. University of Memphis, 2021.
- [211] S. M. Babu, A. J. Lakshmi, B. T. Rao. “A study on cloud based Internet of Things: CloudIoT”. In: *2015 Global Conference on Communication Technologies (GCCT)*. 2015, pp. 60–65. DOI: [10.1109/GCCT.2015.7342624](https://doi.org/10.1109/GCCT.2015.7342624).
- [212] S. P. Mulamoottil and T. Vigneswaran. “A double-layered fully automated insomnia identification model employing synthetic data generation using MCSA and CTGAN with single-channel EEG signals”. In: *Scientific Reports* 14.1 (2024), p. 23427. DOI: [10.1038/s41598-024-74706-9](https://doi.org/10.1038/s41598-024-74706-9).
- [213] S. Prabakaran, M. Wengenroth, H. E. Lockstone, K. Lilley, F. M. Leweke, and S. Bahn. “2-D DIGE analysis of liver and red blood cells provides further evidence for oxidative stress in schizophrenia”. In: *Journal of Proteome Research* 6 (2007), pp. 141–149. DOI: [10.1021/pr060308a](https://doi.org/10.1021/pr060308a).
- [214] S. Sandoval, V. Berisha, R. L. Utianski, J. M. Liss, and A. Spanias. “Automatic assessment of vowel space area”. In: *The Journal of the Acoustical Society of America* 134.5 (2013), EL477–EL483. DOI: [10.1121/1.4826150](https://doi.org/10.1121/1.4826150).
- [215] S. Spijker, J. S. Van Zanten, S. De Jong, B. W. Penninx, R. van Dyck, F. G. Zitman, et al. “Stimulated gene expression profiles as a bloodmarker of major depressive disorder”. In: *Biological Psychiatry* 68.2 (2010), pp. 179–186. DOI: [10.1016/j.biopsych.2010.03.017](https://doi.org/10.1016/j.biopsych.2010.03.017).

- [216] S.-H. Sung, S. Hong, J.-M. Kim, D.-Y. Kang, H. Park, and S. Kim. “Cognitive impairment classification prediction model using voice signal analysis”. In: *Electronics* 13.18 (2024). ISSN: 2079-9292. DOI: [10.3390/electronics13183644](https://doi.org/10.3390/electronics13183644).
- [217] P. Shaver. “Emotion knowledge: Further exploration of a prototype approach”. In: *Journal of Personality and Social Psychology* 52.6 (1987), pp. 1061–1086.
- [218] Snowflake. *How a data ingestion framework powers large data set usage*. Accessed: 10-04-2025. 2025. URL: <https://www.snowflake.com/guides/how-data-ingestion-framework-powers-large-data-set-usage/>.
- [219] C. Spence. “Comfort food: A review”. In: *International Journal of Gastronomy and Food Science* 9 (2017), pp. 105–109. ISSN: 1878-450X. DOI: [10.1016/j.ijgfs.2017.07.001](https://doi.org/10.1016/j.ijgfs.2017.07.001).
- [220] LibreTexts Statistics. *z-Scores*. <https://stats.libretexts.org/@go/page/50029>. Accessed: 01-06-2025. Oct. 2024. URL: <https://stats.libretexts.org/@go/page/50029>.
- [221] T. Ito and K. Takeda, and F. Itakura. “Analysis and recognition of whispered speech”. In: *Speech Communication* 45.2 (2005), pp. 139–152. ISSN: 0167-6393. DOI: <https://doi.org/10.1016/j.specom.2003.10.005>.
- [222] T. Penzel, M. Glos, I. Fietze, S. Herberger, and G. Pillar. “Distinguish obstructive and central sleep apnea by portable peripheral arterial tonometry”. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Vol. 2020-July. 2020, pp. 2780–2783. DOI: [10.1109/EMBC44109.2020.9175700](https://doi.org/10.1109/EMBC44109.2020.9175700).
- [223] T. Winger, V. Chellamuthu, D. Guzenko, M. Aloia, S. Barr, S. DeFranco, B. Gorski, F. Mushtaq, G. Garcia-Molina. “Fine tuned personalized machine learning models to detect insomnia risk based on data from a smart bed platform”. In: *Frontiers in Neurology* Volume 15 - 2024 (2024). ISSN: 1664-2295. DOI: [10.3389/fneur.2024.1303978](https://doi.org/10.3389/fneur.2024.1303978).
- [224] Silero team. *Silero VAD: pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier*. <https://github.com/snakers4/silero-vad>. 2024.
- [225] The editors of Encyclopaedia Britannica. *Hypothetico-deductive method*. <https://www.britannica.com/science/hypothetico-deductive-method>. Encyclopedia Britannica. Mar. 2020.
- [226] U. Yadav, A. Sharma, and D. Patil. “Review of automated depression detection: Social posts, audio and video, open challenges and future direction”. In: *Concurrency and Computation: Practice and Experience* 35 (2022). DOI: [10.1002/cpe.7407](https://doi.org/10.1002/cpe.7407).
- [227] V. Saran, R. Kumar, G. Kumar, K. Chokalingam, M. Rawooth, and G. Parchani. “Validation of dozee, a ballistocardiography-based device, for contactless and continuous heart rate and respiratory rate measurement”. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Vol. 2022-July. 2022. DOI: [10.1109/EMBC48229.2022.9871007](https://doi.org/10.1109/EMBC48229.2022.9871007).
- [228] V. Velepucha and P. Flores. “A survey on microservices architecture: Principles, patterns and migration challenges”. In: *IEEE Access* 11 (2023), pp. 88339–88358. DOI: [10.1109/ACCESS.2023.3305687](https://doi.org/10.1109/ACCESS.2023.3305687).

- [229] Vandana, N. Marriwala, and D. Chaudhary. “A hybrid model for depression detection using deep learning”. In: *Measurement: Sensors* 25 (2023), p. 100587. ISSN: 2665-9174. DOI: <https://doi.org/10.1016/j.measen.2022.100587>.
- [230] W. A. Sanadi, W. A. Hasan. *Acoustic analysis of speech based on power spectral density features in detecting suicidal risk among female patients*. 2011.
- [231] W.-C. Lin and C. Busso. “An efficient temporal modeling approach for speech emotion recognition by mapping varied duration sentences into fixed number of chunks”. In: *Interspeech 2020*. 2020, pp. 2322–2326. DOI: [10.21437/Interspeech.2020-2636](https://doi.org/10.21437/Interspeech.2020-2636).
- [232] X. Huang, F. Wang, Y. Gao, et al. “Depression recognition using voice-based pre-training model”. In: *Scientific Reports* 14 (2024), p. 12734. DOI: [10.1038/s41598-024-63556-0](https://doi.org/10.1038/s41598-024-63556-0).
- [233] X. J. Guo, P. Wu, X. Jia, Y. M. Dong, C. M. Zhao, N. N. Chen, Z. Y. Zhang, Y. T. Miao, K. M. Yun, C. R. Gao, and Y. Ren. “Mapping the structure of depression biomarker research: A bibliometric analysis”. In: *Frontiers in Psychiatry* 13 (2022), p. 943996. DOI: [10.3389/fpsyt.2022.943996](https://doi.org/10.3389/fpsyt.2022.943996).
- [234] X. Kong, Y. Yao, C. Wang, Y. Wang, J. Teng, and X. Qi. “Automatic identification of depression using facial images with deep convolutional neural network”. In: *Medical Science Monitor, International Medical Journal of Experimental and Clinical Research* 28 (2022), e936409. DOI: [10.12659/MSM.936409](https://doi.org/10.12659/MSM.936409).
- [235] X. Liu and K. Jiang. “Why is diagnosing MDD challenging?” In: *Shanghai Archives of Psychiatry* 28.6 (2016), pp. 343–345. DOI: [10.11919/j.issn.1002-0829.216073](https://doi.org/10.11919/j.issn.1002-0829.216073).
- [236] X. Zhang, Z. Zhang, W. Diao, C. Zhou, Y. Song, R. Wang, X. Luo, and G. Liu. “Early-diagnosis of major depressive disorder: From biomarkers to point-of-care testing”. In: *TrAC Trends in Analytical Chemistry* 159 (2023), p. 116904. ISSN: 0165-9936. DOI: <https://doi.org/10.1016/j.trac.2022.116904>.
- [237] Y. Adi, J. Keshet, O. Dmitrieva, M. Goldrick. “Automatic measurement of voice onset time and prevoicing using recurrent neural networks”. In: 2016, pp. 3152–3155. DOI: [10.21437/Interspeech.2016-893](https://doi.org/10.21437/Interspeech.2016-893).
- [238] Y. Aina and J. L. Susman. “Understanding comorbidity with depression and anxiety disorders”. In: *Journal of the American Osteopathic Association* 106.5 Suppl 2 (2006), S9–S14.
- [239] Y. Di, E. Rahmani, J. Mefford, et al. “Unraveling the associations between voice pitch and major depressive disorder: a multisite genetic study”. In: *Molecular Psychiatry* (2024). DOI: [10.1038/s41380-024-02877-y](https://doi.org/10.1038/s41380-024-02877-y).
- [240] Y. J. Lee, J. Palmer, A. C. Cline, H. Lee. “Factors Influencing the Health Care Utilization Among People With Depression and/or Anxiety Symptoms”. In: *Journal of the American Psychiatric Nurses Association* 30.5 (2024), pp. 858–867. DOI: [10.1177/10783903231197655](https://doi.org/10.1177/10783903231197655).
- [241] Y. Liu, X. Li, M. Wang, J. Bi, S. Lin, Q. Wang, Y. Yu, J. Ye, and Y. Zheng. “Multimodal depression recognition and analysis: Facial expression and body posture changes via emotional stimuli”. In: *Journal of Affective Disorders* 381 (2025), pp. 44–54. ISSN: 0165-0327. DOI: <https://doi.org/10.1016/j.jad.2025.03.155>.

- [242] Y. Wu, et al. “Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: A systematic review and individual participant data meta-analysis”. In: *Psychological Medicine* 50.8 (2020), pp. 1368–1380. DOI: [10.1017/S0033291719001314](https://doi.org/10.1017/S0033291719001314).
- [243] Y. Yang, C. Fairbairn, and J. F. Cohn. “Detecting depression severity from vocal prosody”. In: *IEEE Transactions on Affective Computing* 4.2 (2013), pp. 142–150. DOI: [10.1109/T-AFFC.2012.38](https://doi.org/10.1109/T-AFFC.2012.38).
- [244] Y. Zhao, Z. Liang, J. Du, L. Zhang, C. Liu, and L. Zhao. “Multi-head attention-based long short-term memory for depression detection from speech”. In: *Frontiers in Neurorobotics* Volume 15 - 2021 (2021). ISSN: 1662-5218. DOI: [10.3389/fnbot.2021.684037](https://doi.org/10.3389/fnbot.2021.684037).
- [245] Y.-G. Wu, Z.-E. Jin, and C.-L. Lin. “Emotion detection by speech and voice tone”. In: 2021.
- [246] T. Yingthawornsuk. “Acoustic analysis of vocal output characteristics for suicidal risk assessment”. Ph.D. Thesis. Vanderbilt University, 2007.
- [247] B. J. Yoon. “Hidden Markov Models and their applications in biological sequence analysis”. In: *Current Genomics* 10.6 (2009), pp. 402–415. DOI: [10.2174/138920209789177575](https://doi.org/10.2174/138920209789177575).
- [248] Z. C. Lipton. “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.” In: *Queue* 16.3 (2018), pp. 31–57.
- [249] Z. Heydarifard and D. Krasikova. “Voice and insomnia: A daily study of underlying affective and cognitive mechanisms”. In: *Academy of Management Proceedings*. Vol. 2021. 1. Academy of Management Briarcliff Manor, NY 10510. 2021, p. 15460.
- [250] Z. Wang, S. B. Ho, and E. Cambria. “A review of emotion sensing: categorization models and algorithms”. In: *Multimedia Tools and Applications* 79 (2020), pp. 35553–35582. DOI: [10.1007/s11042-019-08328-z](https://doi.org/10.1007/s11042-019-08328-z).

Appendix A

Questionnaires



Personal Health Questionnaire Depression Scale (PHQ-8)

Over the **last 2 weeks**, how often have you been bothered by any of the following problems? (*circle one number on each line*)

How often during the past 2 weeks were you bothered by...	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself, or that you are a failure, or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed. Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3

Scoring

If two consecutive numbers are circled, score the higher (more distress) number. If the numbers are not consecutive, do not score the item. Score is the sum of the 8 items. If more than 1 item missing, set the value of the scale to missing. A score of 10 or greater is considered major depression, 20 or more is severe major depression.

Characteristics

Tested on 1165 subjects with chronic conditions.

No. of items	Observed Range	Mean	Standard Deviation	Internal Consistency Reliability	Test-Retest Reliability
8	0-24	6.63	5.52	.86	NA

This scale is free to use without permission

1

Figure A.1: PHQ-8 questionnaire (1/2) [110].

Source of Psychometric Data

U.S. National Chronic Disease Self-Management Study. Study described in Ory MG, Ahn S, Jiang L, et al. National study of chronic disease self-management: six month outcome findings. Journal of Aging and Health. 2013 [in press].

Comments

This is an adaptation of the PHQ-9 scale. Since this scale is self-administered in our studies, question #9, "How often during the past 2 weeks were you bothered by thoughts that you would be better off dead, or of hurting yourself in some way?", was deleted. This scale available in Spanish.

References

Kroenke K, Strine TW, Spritzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord. 2009; 114(1-3):163-73.

Razykov I, Ziegelstein RC, Whooley MA, Thombs BD. The PHQ-9 versus the PHQ-8--is item 9 useful for assessing suicide risk in coronary artery disease patients? Data from the Heart and Soul Study. J Psychosom Res. 2012; 73(3):163-168.

This scale is free to use without permission

Self-Management Resource Center
smrc@selfmanagementresource.com
<https://selfmanagementresource.com>

This scale is free to use without permission

2

Figure A.2: PHQ-8 questionnaire (2/2) [110].

PATIENT HEALTH QUESTIONNAIRE (PHQ-9)

ID #: _____ DATE: _____

Over the last 2 weeks, how often have you been
bothered by any of the following problems?
(use "✓" to indicate your answer)

	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself—or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed. Or the opposite — being so figety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead, or of hurting yourself	0	1	2	3

add columns + +

(Healthcare professional: For interpretation of TOTAL, TOTAL:
please refer to accompanying scoring card).

10. If you checked off <i>any problems</i> , how <i>difficult</i> have these problems made it for you to do your work, take care of things at home, or get along with other people?	Not difficult at all	_____
	Somewhat difficult	_____
	Very difficult	_____
	Extremely difficult	_____

Copyright © 1999 Pfizer Inc. All rights reserved. Reproduced with permission. PRIME-MD® is a trademark of Pfizer Inc.
A2663B 10-04-2005

Figure A.3: PHQ-9 questionnaire (1/2) [186].

PHQ-9 Patient Depression Questionnaire

For initial diagnosis:

1. Patient completes PHQ-9 Quick Depression Assessment.
2. If there are at least 4 ✓s in the shaded section (including Questions #1 and #2), consider a depressive disorder. Add score to determine severity.

Consider Major Depressive Disorder

- if there are at least 5 ✓s in the shaded section (one of which corresponds to Question #1 or #2)

Consider Other Depressive Disorder

- if there are 2-4 ✓s in the shaded section (one of which corresponds to Question #1 or #2)

Note: Since the questionnaire relies on patient self-report, all responses should be verified by the clinician, and a definitive diagnosis is made on clinical grounds taking into account how well the patient understood the questionnaire, as well as other relevant information from the patient.

Diagnoses of Major Depressive Disorder or Other Depressive Disorder also require impairment of social, occupational, or other important areas of functioning (Question #10) and ruling out normal bereavement, a history of a Manic Episode (Bipolar Disorder), and a physical disorder, medication, or other drug as the biological cause of the depressive symptoms.

To monitor severity over time for newly diagnosed patients or patients in current treatment for depression:

1. Patients may complete questionnaires at baseline and at regular intervals (eg, every 2 weeks) at home and bring them in at their next appointment for scoring or they may complete the questionnaire during each scheduled appointment.
2. Add up ✓s by column. For every ✓: Several days = 1 More than half the days = 2 Nearly every day = 3
3. Add together column scores to get a TOTAL score.
4. Refer to the accompanying **PHQ-9 Scoring Box** to interpret the TOTAL score.
5. Results may be included in patient files to assist you in setting up a treatment goal, determining degree of response, as well as guiding treatment intervention.

Scoring: add up all checked boxes on PHQ-9

For every ✓ Not at all = 0; Several days = 1;
More than half the days = 2; Nearly every day = 3

Interpretation of Total Score

Total Score	Depression Severity
1-4	Minimal depression
5-9	Mild depression
10-14	Moderate depression
15-19	Moderately severe depression
20-27	Severe depression

PHQ9 Copyright © Pfizer Inc. All rights reserved. Reproduced with permission. PRIME-MD ® is a trademark of Pfizer Inc.

A2662B 10-04-2005

Figure A.4: PHQ-9 questionnaire (2/2) [186].

Appendix B

Prototype

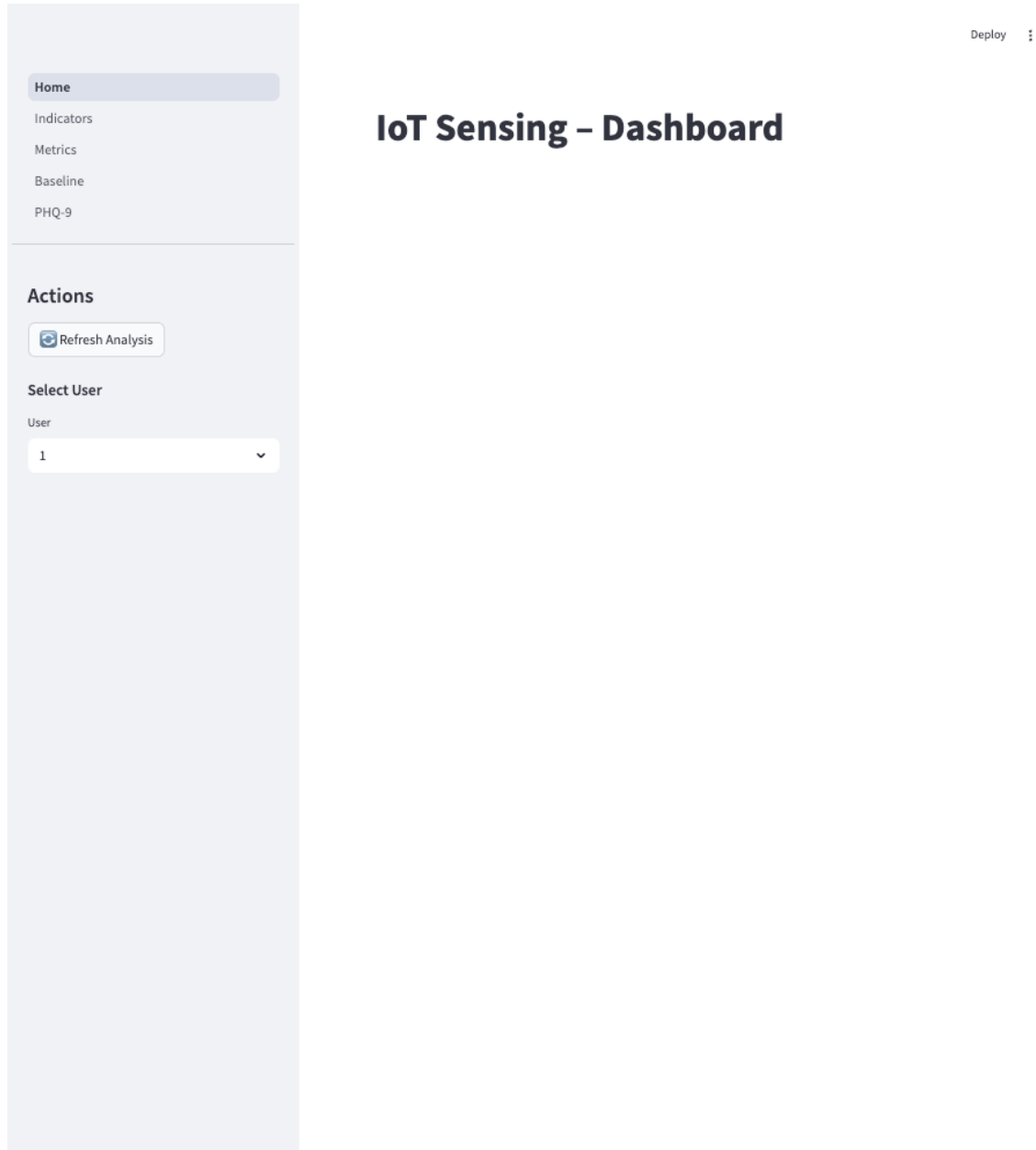


Figure B.1: Prototype screenshot (1/8) – Home [128].

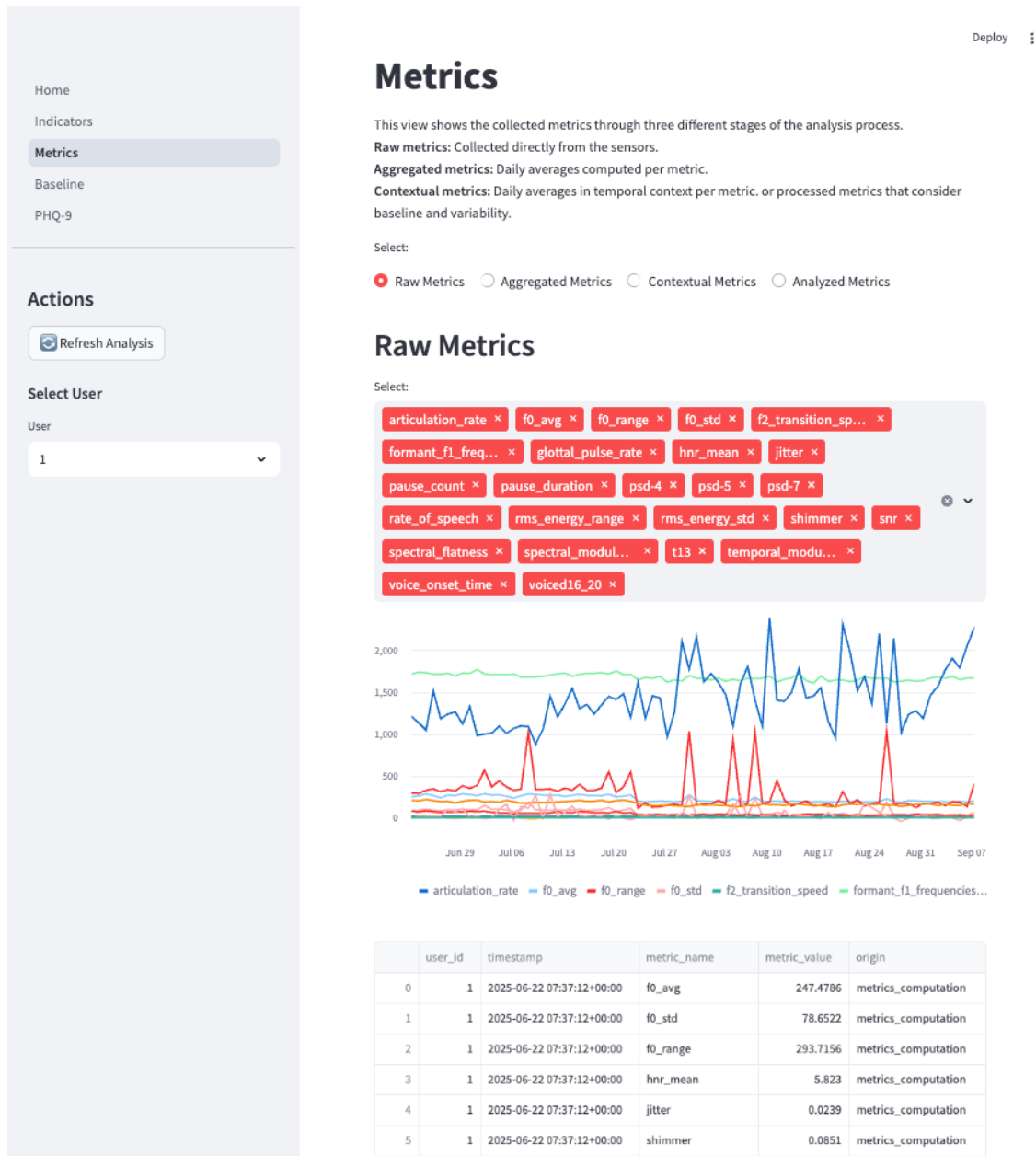


Figure B.2: Prototype screenshot (2/8) – Raw Metrics [128].

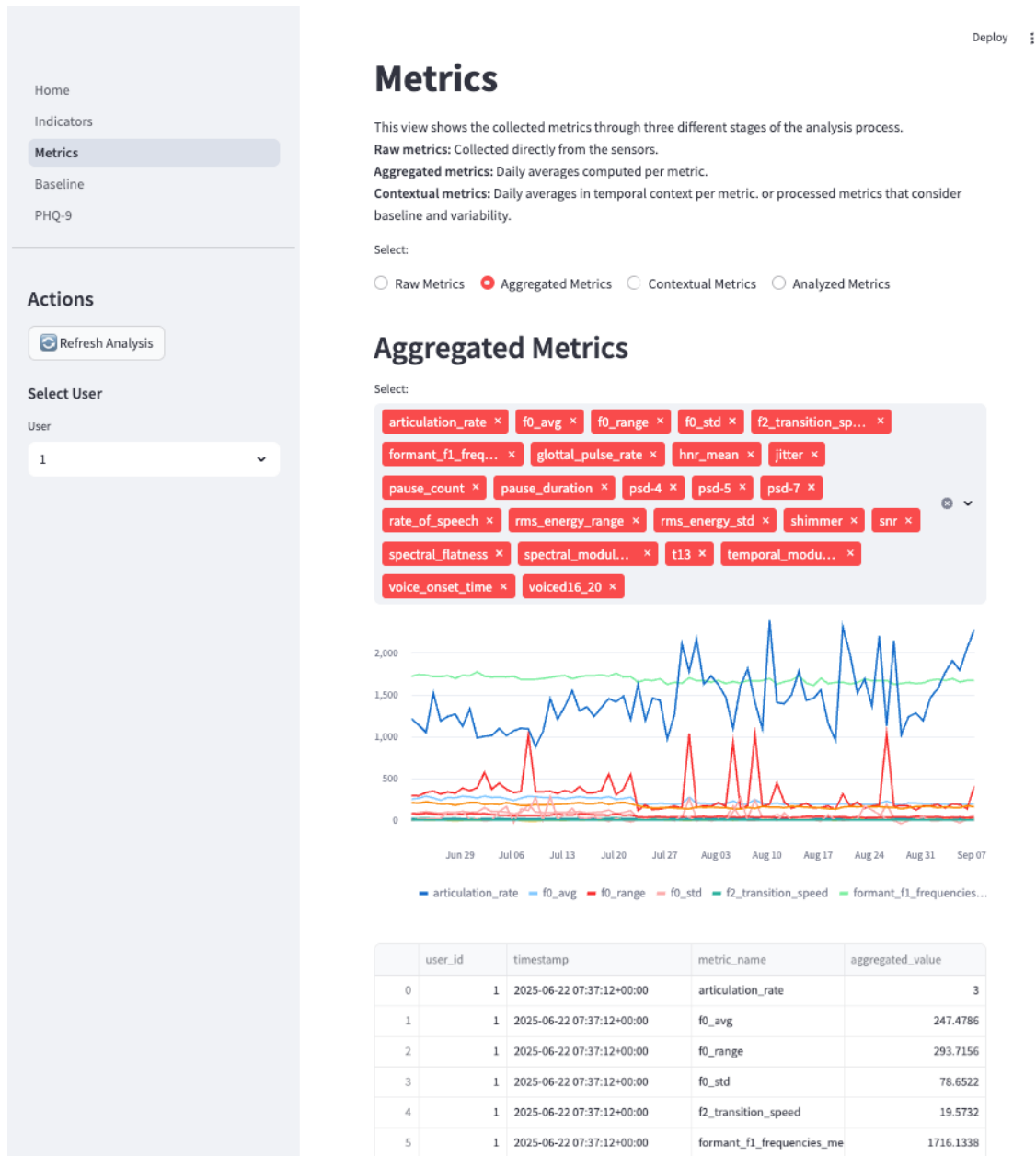


Figure B.3: Prototype screenshot (3/8) – Aggregated Metrics [128].

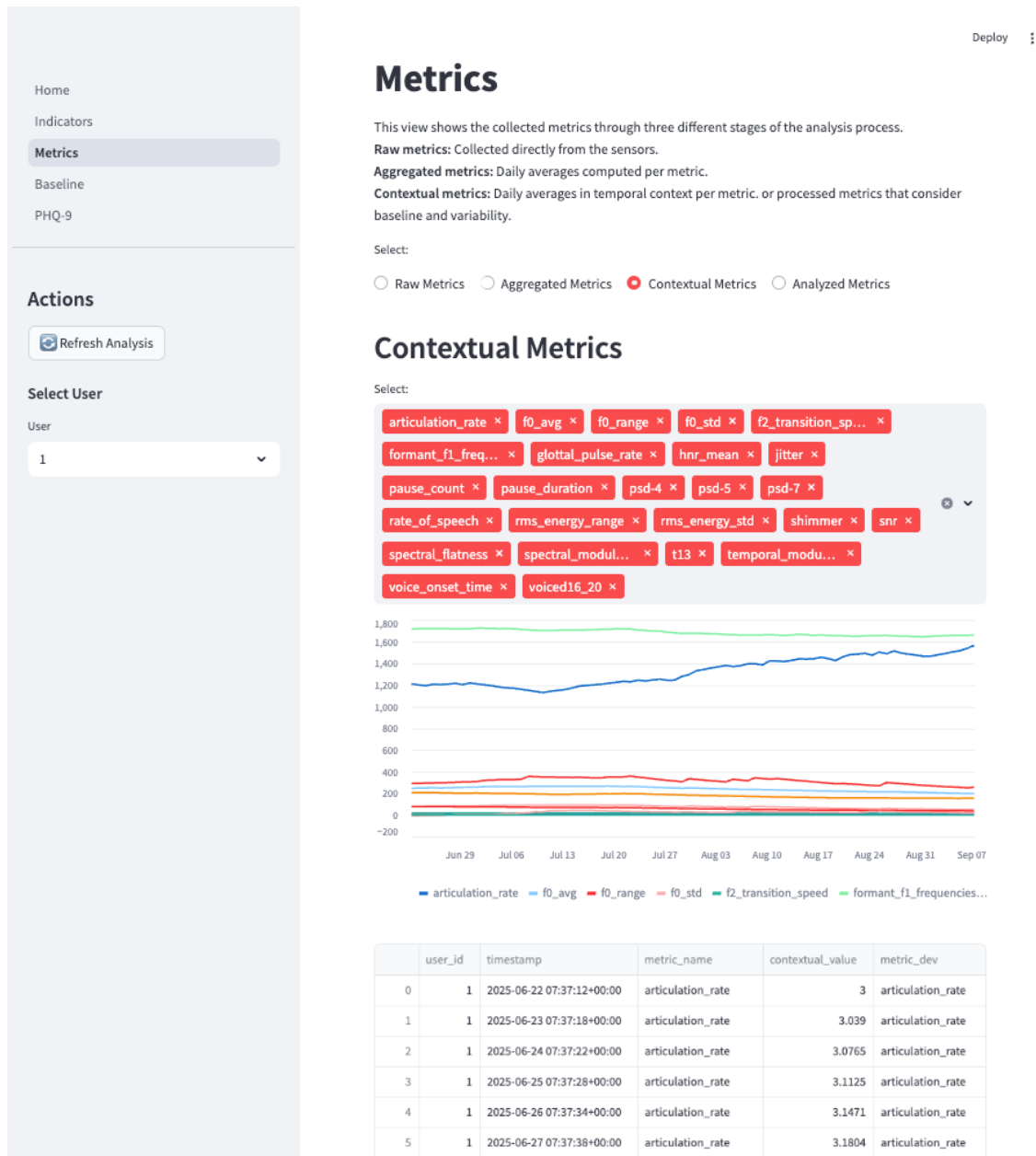


Figure B.4: Prototype screenshot (4/8) – Contextual Metrics [128].

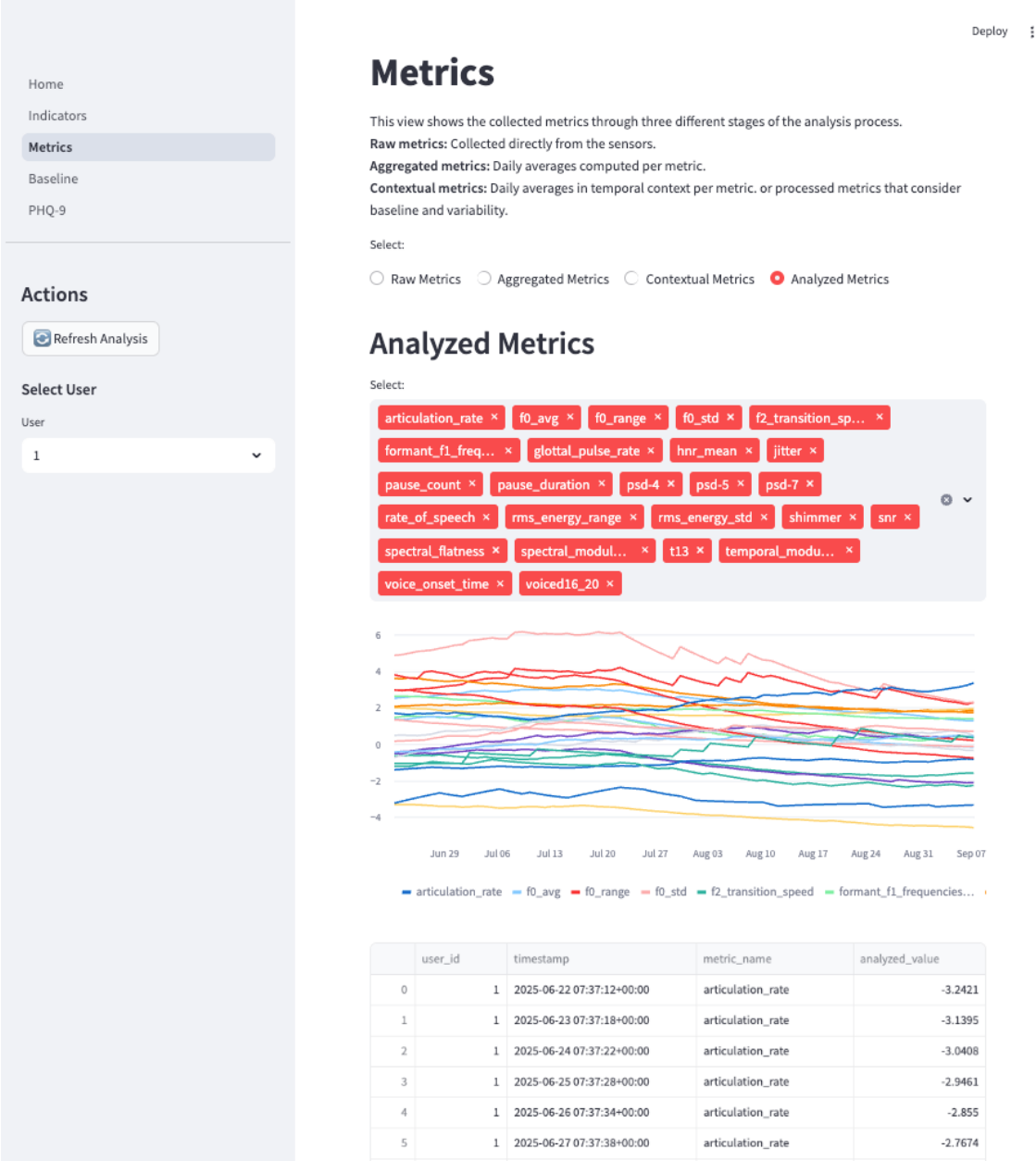


Figure B.5: Prototype screenshot (5/8) – Analyzed Metrics [128].

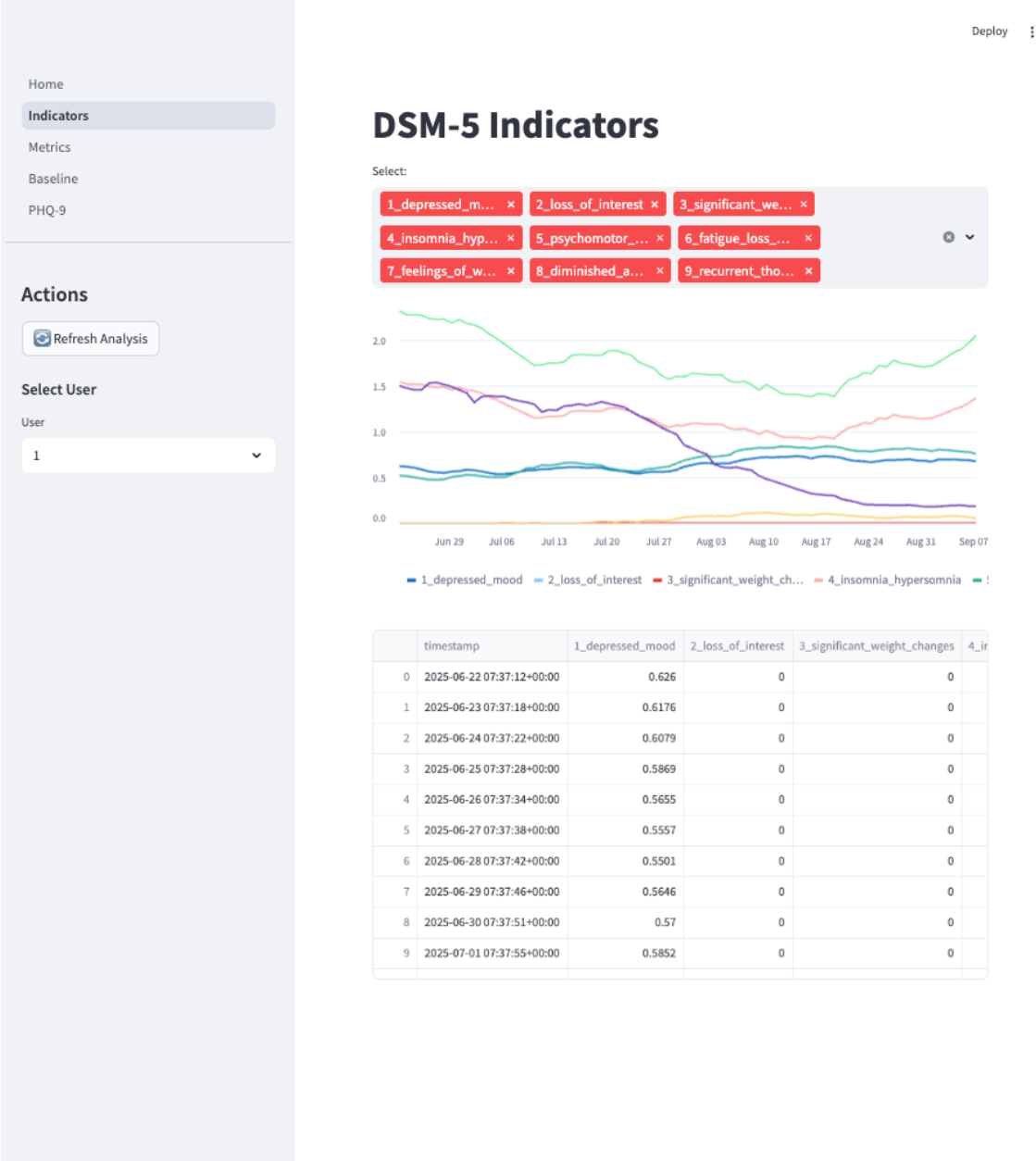


Figure B.6: Prototype screenshot (6/8) – DSM-5 Indicator Scores [128].

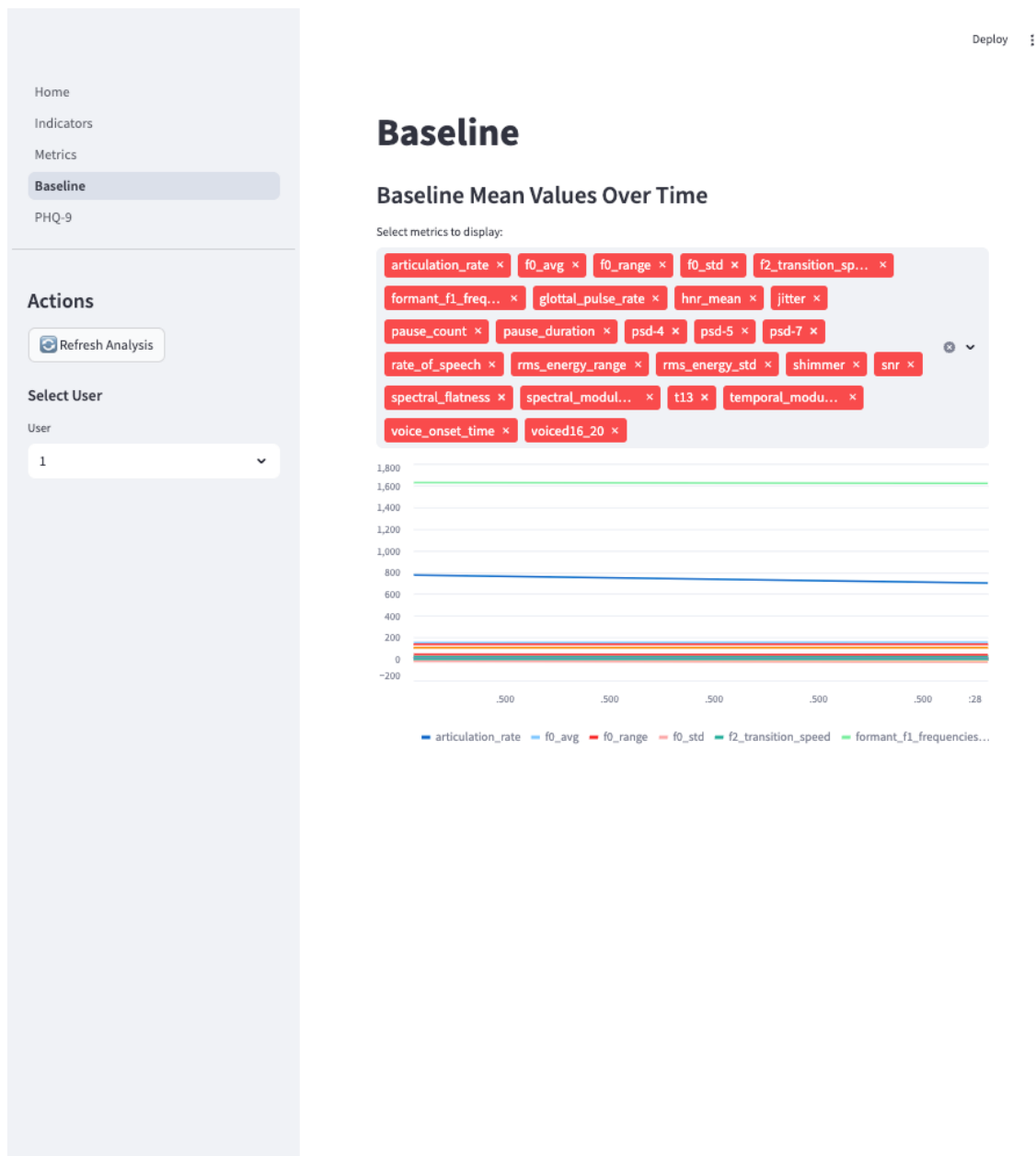


Figure B.7: Prototype screenshot (7/8) – Baseline [128].

Deploy

Home

Indicators

Metrics

Baseline

PHQ-9

Actions

Refresh Analysis

Select User

User

1

PHQ-9 Questionnaire

Over the last 2 weeks, how often have you been bothered by any of the following problems?

Little interest or pleasure in doing things

Not at all (0)

Feeling down, depressed, or hopeless

Not at all (0)

Trouble falling or staying asleep, or sleeping too much

Not at all (0)

Feeling tired or having little energy

Not at all (0)

Poor appetite or overeating

Not at all (0)

Feeling bad about yourself – or that you are a failure or have let yourself or your family down

Not at all (0)

Trouble concentrating on things, such as reading the newspaper or watching television

Not at all (0)

Moving or speaking so slowly that other people could have noticed. Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual

Not at all (0)

Thoughts that you would be better off dead, or of hurting yourself

Not at all (0)

Total Score: 0

If you checked off any problems, how difficult have these problems made it for you to do your work, take care of things at home, or get along with other people?

Not difficult at all

Submit

Figure B.8: Prototype screenshot (8/8) – PHQ-9 Questionnaire [128].

Appendix C

Evaluation

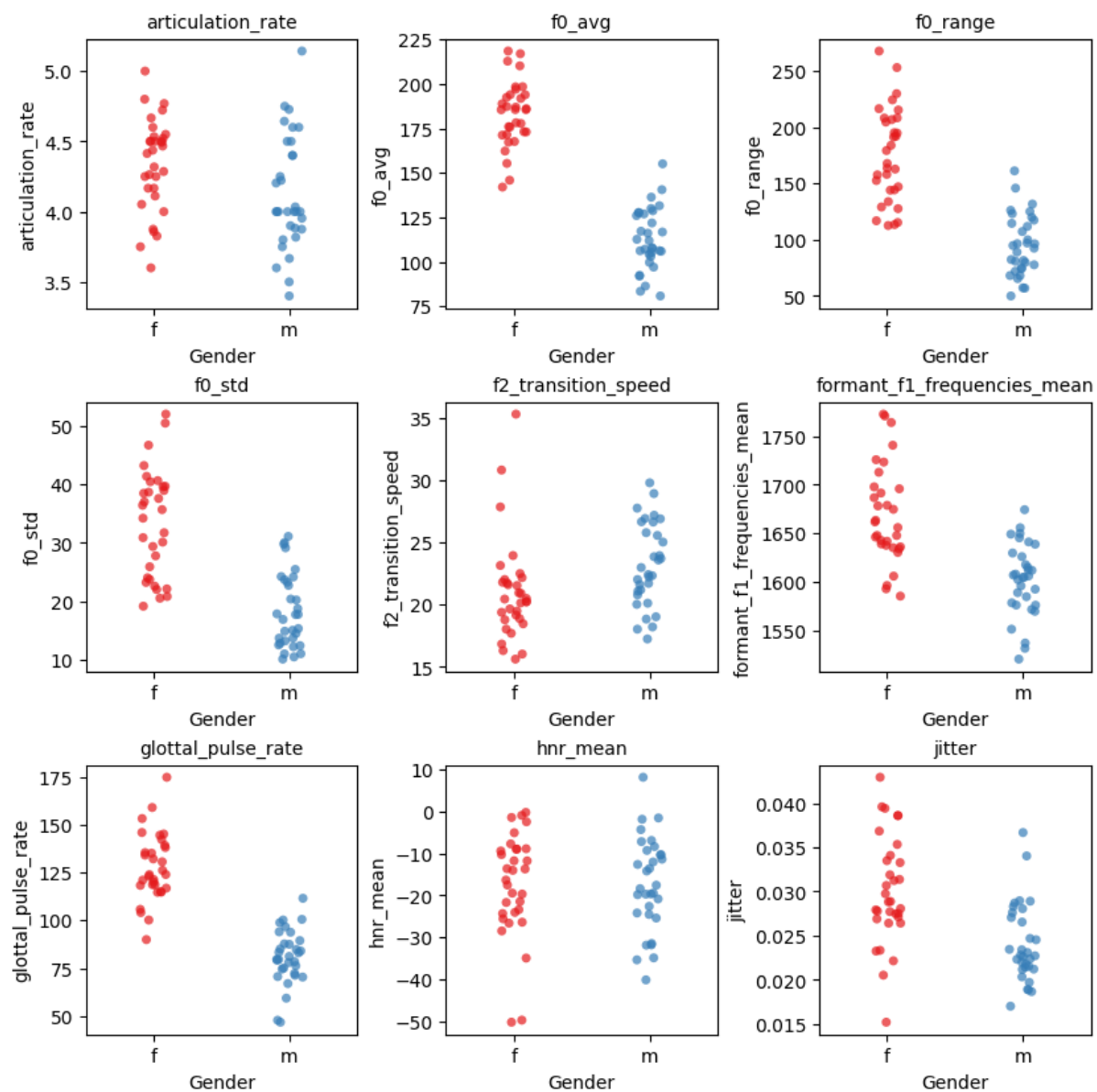


Figure C.1: Acoustic voice feature sample distribution – scatter plot (1/3) [128].

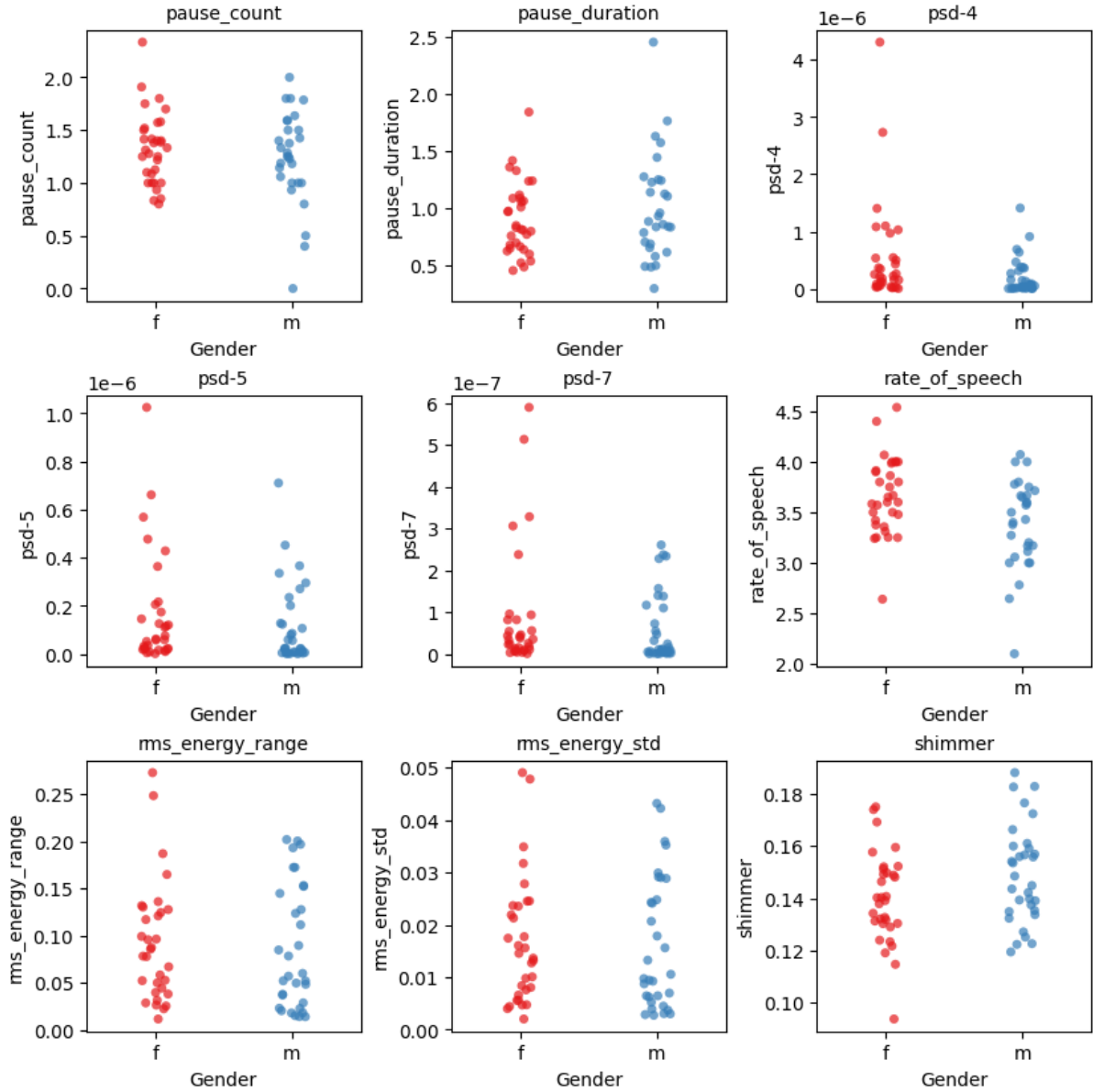


Figure C.2: Acoustic voice feature sample distribution – scatter plot (2/3) [128].

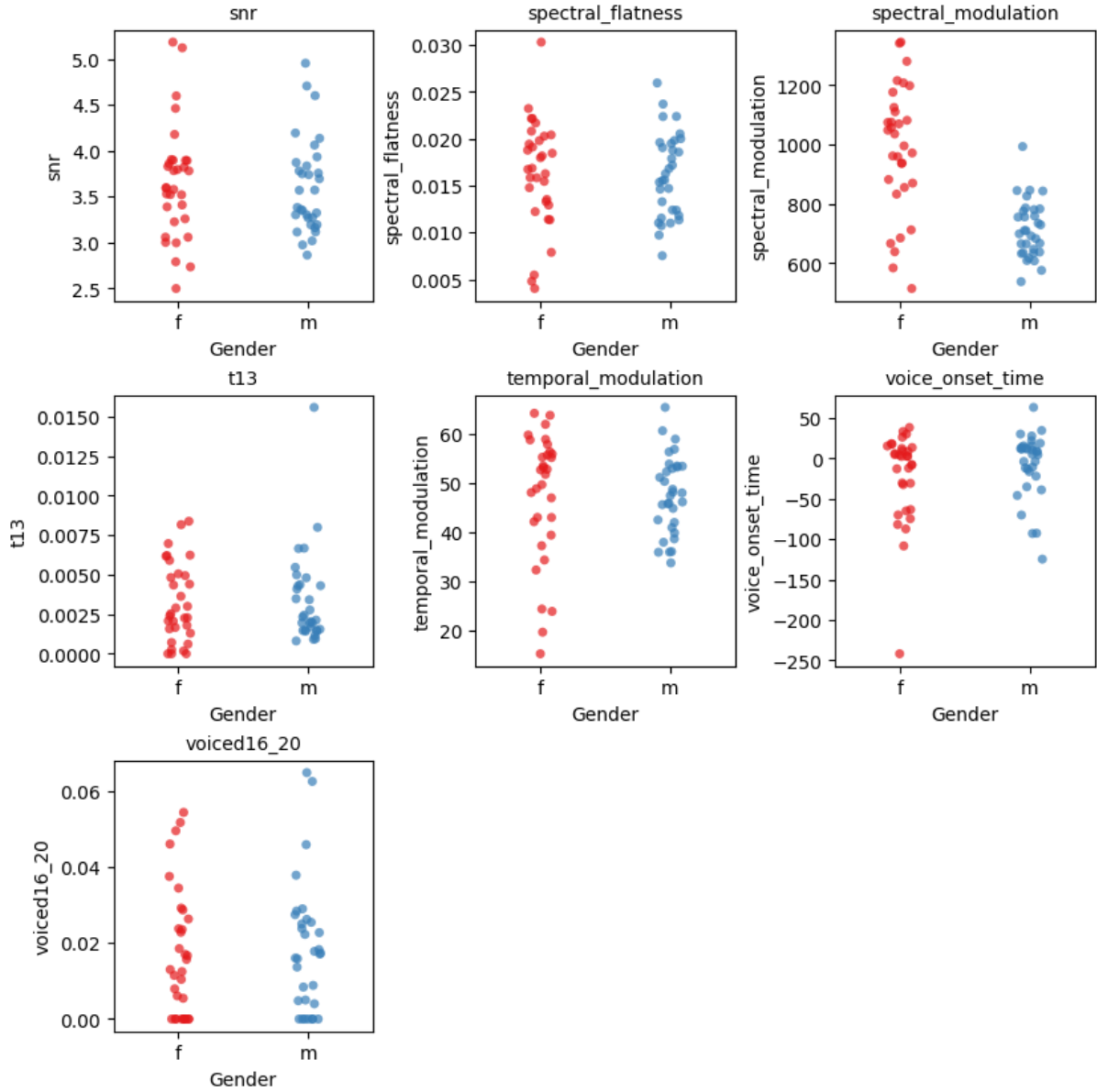


Figure C.3: Acoustic voice feature sample distribution – scatter plot (3/3) [128].

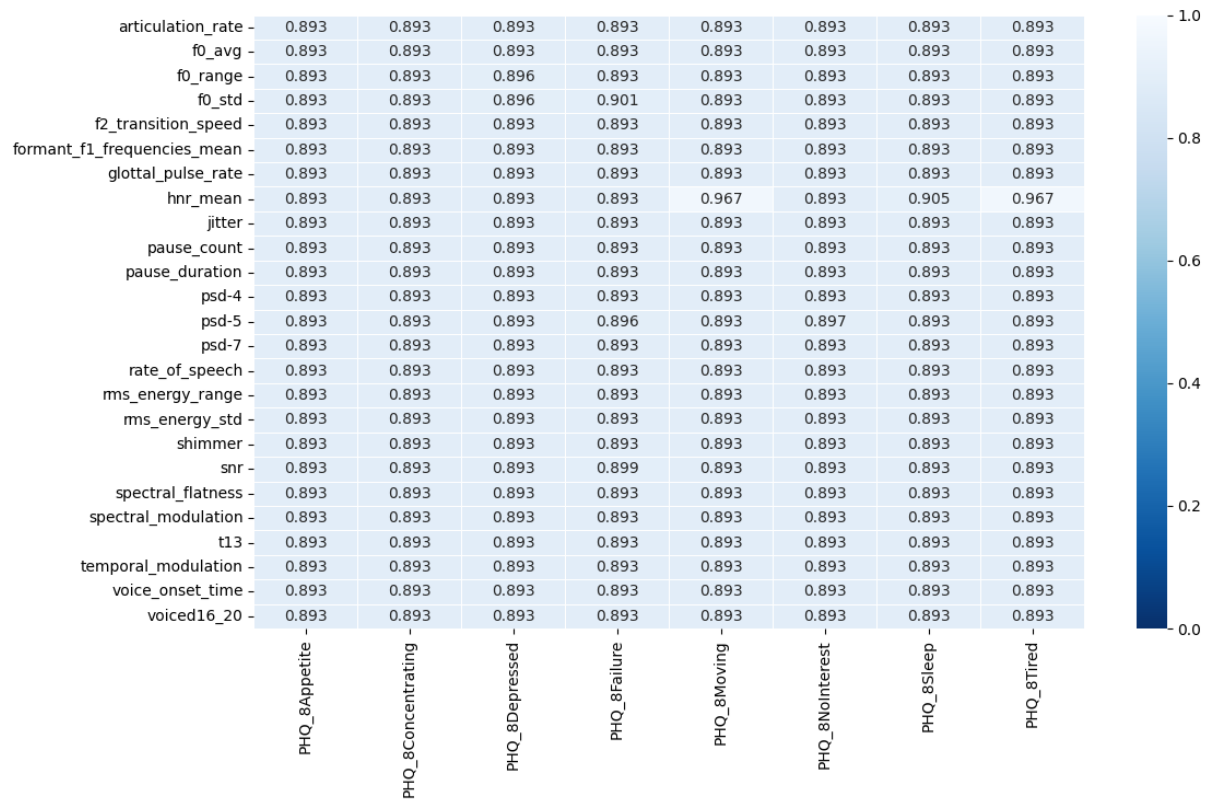


Figure C.4: Adjusted Pearson p-values of feature-to-indicator pairings [128].

Appendix D

GitHub Repository

GitHub – Repository [128]: <https://github.com/jonaslanzlinger/depression-detection>.

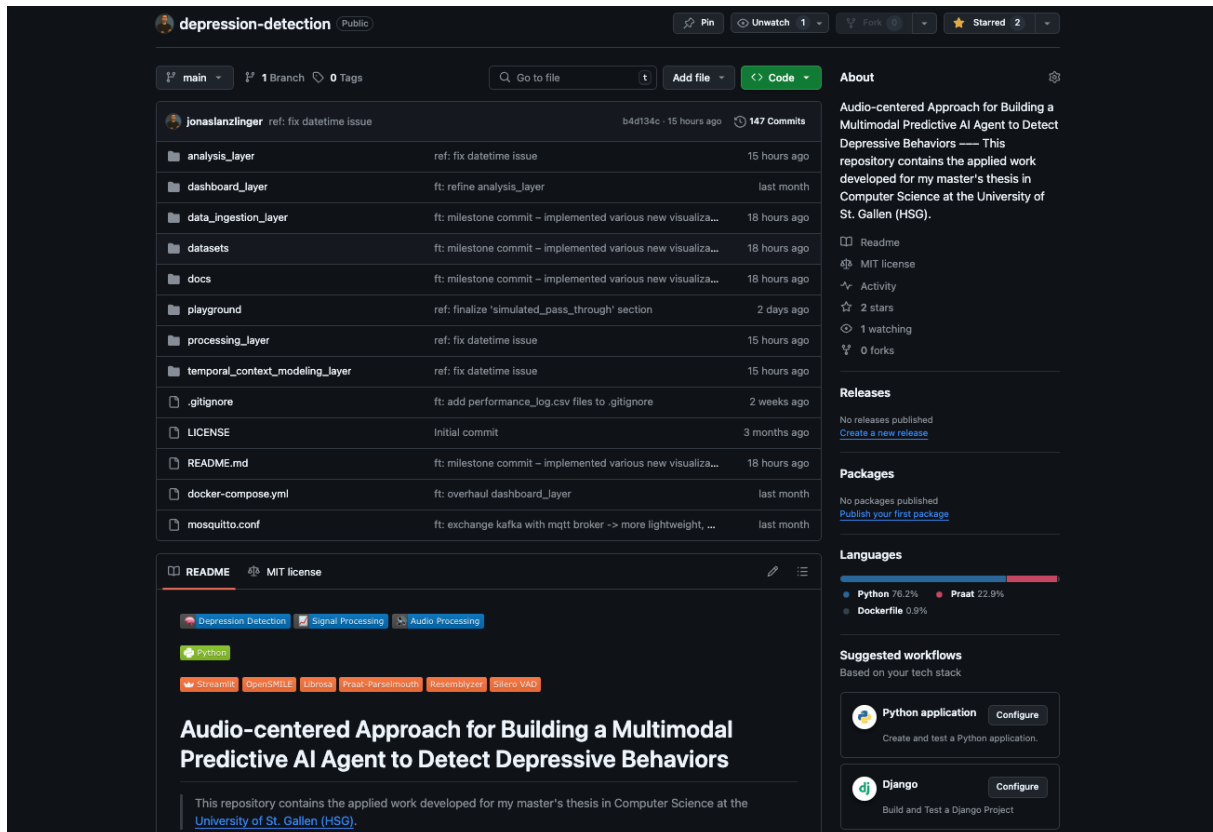


Figure D.1: GitHub repository *depression-detection* [128].

Repository Overview

This repository is structured as follows:

- **analysis_layer**: Reflects the *Analysis Layer* component on the data pipeline and is responsible for quantifying the feature-to-indicator mapping using the concrete employed *config.json* and the respective user baseline (initially: *population_baseline.json*).
- **dashboard_layer**: Reflects the *Dashboard Layer* component on the data pipeline and is responsible for the user-system interaction and providing meaningful visualizations of the system output.

-
- **data_ingestion_layer**: Reflects the *Data Ingestion Layer* component on the data pipeline and is responsible for setting up the sensor types and streaming the data to the system core.
 - */framework/* – contains interfaces for uniform implementation of new sensor modalities. (user may not extend)
 - */implementation/* – contains implemented sensor modalities. (user may extend)
 - **datasets**: Contains all data files that are possible to share, such as the Toronto Emotional Speech Set (TESS). Due to license agreements, the DAIC-WOZ speech set has to be kept hidden (in */ex_files/*).
 - **docs**: Contains various scripts that have been used for creating visualizations or the evaluation of the master's thesis and miscellaneous additional resources such as data files and log-files.
 - **playground**: Experimental content that I like to keep in the repository for personal use.
 - **processing_layer**: Reflects the *Processing Layer* component on the data pipeline and employs two types of service groups: (1) metrics computation services, and (2) user profiling services.
 - **temporal_context_modeling_layer**: Reflects the *Temporal Context Modeling Layer* component on the data pipeline and is responsible for aggregating the raw metrics as well as contextualize them to make the sensed values more rigid and having them reflect the temporal notion as described in DSM-5.

Declaration of Authorship

‘I hereby declare,

- that I have written this thesis independently;
- that I have written the articles and contributions using only the aids listed in the index;
- that all parts of the thesis produced with the help of aids (incl. AI-Tools) have been precisely declared;
- that I have mentioned all sources used and cited them correctly according to established academic citation rules; this also includes the comprehensible disclosure of all personal publications;
- that I have acquired all immaterial rights to any materials I may have used, such as images or graphics, or that these materials were originally created by myself;
- that I am aware of the legal provisions regarding the publication and dissemination of parts or the entire thesis and that I comply with them accordingly;
- that I am aware that the University will prosecute a violation of this Declaration of Authorship and that disciplinary as well as criminal consequences may result, which may lead to expulsion from the University or to the withdrawal of my title.’

By submitting this thesis, I confirm through my conclusive action that I am submitting the statutory declaration, that I have read and understood it, and that it is true.

St. Gallen, June 26, 2025

Signature: Jonas Länzlinger

Declaration of Auxiliary Aids

The creation of this master's thesis involved the use of third-party software, including AI-based tools. I declare all uses of third-party software in this work that contributed to the implementation, text, or the research in any way. Further, by the University of St. Gallen's decree II.B.1.20 section 5.3, I explicitly declare the following uses:

Aid	Usage	Affected Parts
GPT-4o	Coding assistance	Code
GPT-4o, Google Translate, Writefull	Spell and grammar checking, translation, and rewordings	Complete paper
GPT-4o	Brainstorming	Complete work