



**Machine Learning Explainability on Multi-Modal Data
using Ecological Momentary Assessments in the Medical Domain**

**Erklärbarkeit von maschinellem Lernen
unter Verwendung multi-modaler Daten
und Ecological Momentary Assessments im medizinischen Sektor**

Doctoral thesis for a doctoral degree
at the Graduate School of Life Sciences
Julius-Maximilians-Universität Würzburg
Section Clinical Sciences

submitted by

Johannes Allgaier

from Aalen, Germany

Würzburg, 2023



Submitted on:

Office stamp

Members of the Thesis Committee:

Chairperson:	Prof. Dr. Manfred Gessler
Primary Supervisor:	Prof. Dr. Rüdiger Pryss
Supervisor (Second):	Prof. Dr. Karandeep Singh
Supervisor (Third):	Prof. Dr. Winfried Schlee
Supervisor (Fourth):	Prof. Dr. Thomas Dandekar

Date of Public Defense:

Date of Receipt of Certificates:



**Machine Learning Explainability
on Multi-Modal Data
using Ecological Momentary Assessments
in the Medical Domain**

Johannes Allgaier

Supervised by
Prof. Dr. Rüdiger Pryss
Prof. Dr. Karandeep Singh
Prof. Dr. Winfried Schlee
Prof. Dr. Thomas Dandekar

Institute for Clinical Epidemiology and Biometry

Faculty of Medicine

Section Clinical Sciences

Julius-Maximilians-Universität Würzburg

September 2023

*A dissertation submitted in partial fulfilment of the requirements for the
degree of Doctor rerum naturalium.*

Meinen Eltern

für die Ermöglichung meiner Ausbildung.

Acknowledgements

My doctoral advisor, **Rüdiger Pryss**, was always approachable throughout my entire dissertation, and often, 5-minute discussions turned into hour-long conversations. He allowed me a lot of time for my own research ideas and projects alongside the dissertation. He never denied any of my requests, fully supported all of my ideas, and usually completed reviews of my work at night and on weekends. I am very grateful to have had you as my doctoral advisor; someone like you is exceptional. I look forward to working on more projects with you in the future.

I also extend my gratitude to other colleagues who have guided me with their experience and insights, including **Winfried Schlee and Johannes Schobel**.

I want to give special thanks to **Karandeep Singh** for his willingness to accommodate me in his lab in the USA and for taking me to the symposium in Puerto Rico. It was an amazing experience, Karandeep, thank you.

I thank my team at ATR Software for the professional exchange on Machine Learning and the shared lunch breaks, including **Burkhardt Hoppenstedt, Arthur Ulmer, Christian Fuchs, Julian Henning, Tobias Hofmann, and Alexander Treß**.

Lastly, I thank **Felizitas Eichner**. She especially supported me during frustrating phases, listened to me, and helped me clear my mind after work. I am thankful to you for that.

Summary

Introduction. Mobile health (mHealth) integrates mobile devices into healthcare, enabling remote monitoring, data collection, and personalized interventions. Machine Learning (ML), a subfield of Artificial Intelligence (AI), can use mHealth data to confirm or extend domain knowledge by finding associations within the data, i.e., with the goal of improving healthcare decisions. In this work, two data collection techniques were used for mHealth data fed into ML systems: Mobile Crowdsensing (MCS), which is a collaborative data gathering approach, and Ecological Momentary Assessments (EMA), which capture real-time individual experiences within the individual's common environments using questionnaires and sensors. We collected EMA and MCS data on tinnitus and COVID-19. About 15 % of the world's population suffers from tinnitus.

Materials & Methods. This thesis investigates the challenges of ML systems when using MCS and EMA data. It asks: How can ML confirm or broad domain knowledge? Domain knowledge refers to expertise and understanding in a specific field, gained through experience and education. Are ML systems always superior to simple heuristics and if yes, how can one reach explainable AI (XAI) in the presence of mHealth data? An XAI method enables a human to understand why a model makes certain predictions. Finally, which guidelines can be beneficial for the use of ML within the mHealth domain? In tinnitus research, ML discerns gender, temperature, and season-related variations among patients. In the realm of COVID-19, we collaboratively designed a [COVID-19 check app](#) for public education, incorporating EMA data to offer informative feedback on COVID-19-related matters. This thesis uses seven EMA datasets with more than 250,000 assessments. Our analyses revealed a set of challenges: App user over-representation, time gaps, identity ambiguity, and operating system specific rounding errors, among others. Our [systematic review](#) of 450 medical studies assessed prior utilization of XAI methods.

Results. ML models [predict gender](#) and [tinnitus perception](#), validating gender-linked tinnitus disparities. Using season and temperature to predict tinnitus shows the association of these variables with tinnitus. Multiple assessments of one app user can constitute a group. Neglecting these groups in data sets leads to model overfitting. In select instances, heuristics outperform ML models, highlighting the need for domain expert consultation to unveil hidden groups or find simple heuristics.

Conclusion. This thesis suggests guidelines for mHealth related data analyses and improves estimates for ML performance. Close communication with medical domain experts to identify latent user subsets and incremental benefits of ML is essential.

Zusammenfassung

Einleitung. Unter Mobile Health (mHealth) versteht man die Nutzung mobiler Geräte wie Handys zur Unterstützung der Gesundheitsversorgung. So können Ärzt:innen z. B. Gesundheitsinformationen sammeln, die Gesundheit aus der Ferne überwachen, sowie personalisierte Behandlungen anbieten. Man kann maschinelles Lernen (ML) als System nutzen, um aus diesen Gesundheitsinformationen zu lernen. Das ML-System versucht, Muster in den mHealth Daten zu finden, um Ärzt:innen zu helfen, bessere Entscheidungen zu treffen. Zur Datensammlung wurden zwei Methoden verwendet: Einerseits trugen zahlreiche Personen zur Sammlung von umfassenden Informationen mit mobilen Geräten bei (sog. *Mobile Crowdsensing*), zum anderen wurde den Mitwirkenden digitale Fragebögen gesendet und Sensoren wie GPS eingesetzt, um Informationen in einer alltäglichen Umgebung zu erfassen (sog. *Ecological Momentary Assessments*). Diese Arbeit verwendet Daten aus zwei medizinischen Bereichen: Tinnitus und COVID-19. Schätzungen zufolge leidet etwa 15 % der Menschheit an Tinnitus.

Materialien & Methoden. Die Arbeit untersucht, wie ML-Systeme mit mHealth Daten umgehen: Wie können diese Systeme robuster werden oder neue Dinge lernen? Funktionieren die neuen ML-Systeme immer besser als einfache Daumenregeln, und wenn ja, wie können wir sie dazu bringen, zu erklären, warum sie bestimmte Entscheidungen treffen? Welche speziellen Regeln sollte man außerdem befolgen, wenn man ML-Systeme mit mHealth Daten trainiert? Während der COVID-19-Pandemie entwickelten wir eine [App](#), die den Menschen helfen sollte, sich über das Virus zu informieren. Diese App nutzte Daten der Krankheitssymptome der App Nutzer:innen, um Handlungsempfehlungen für das weitere Vorgehen zu geben.

Ergebnisse. ML-Systeme wurden trainiert, um Tinnitus vorherzusagen und wie er mit geschlechtsspezifischen Unterschieden zusammenhängen könnte. Die Verwendung von Faktoren wie Jahreszeit und Temperatur kann helfen, Tinnitus und seine Beziehung zu diesen Faktoren zu verstehen. Wenn wir beim Training nicht berücksichtigen, dass ein App User mehrere Datensätze ausfüllen kann, führt dies zu einer Überanpassung und damit Verschlechterung des ML-Systems. Interessanterweise führen manchmal einfache Regeln zu robusteren und besseren Modellen als komplexe ML-Systeme. Das zeigt, dass es wichtig ist, Experten auf dem Gebiet einzubeziehen, um Überanpassung zu vermeiden oder einfache Regeln zur Vorhersage zu finden.

Fazit. Durch die Betrachtung verschiedener Langzeitdaten konnten wir neue Empfehlungen zur Analyse von mHealth Daten und der Entwicklung von ML-Systemen ableiten. Dabei ist es wichtig, medizinischen Experten mit einzubeziehen, um Überanpassung zu vermeiden und ML-Systeme schrittweise zu verbessern.

Contents

1	Introduction	1
1.1	Explanation of key concepts of this work	2
1.2	Problem statement and motivation	5
1.2.1	What are specific challenges of data in the context of Ecological Momentary Assessments and Mobile Crowdsensing?	6
1.2.2	Why is machine learning beneficial for evaluation?	10
1.2.3	Why is machine learning explainability needed?	13
1.3	Introduction to thesis-related domains	16
1.3.1	Tinnitus	16
1.3.2	Severe Acute Respiratory Syndrome Coronavirus 2	17
2	Materials & Methods	19
2.1	EMA mHealth datasets	19
2.1.1	Excel-Loop	23
2.1.2	The dashboard	25
2.2	Supervised Machine Learning	26
2.2.1	Bias-variance tradeoff and the problem with overfitting	27
2.2.2	Tree-based ML methods	28
2.3	Machine Learning Pipelines	29
2.3.1	CRISP-DM	30
2.3.2	Cross-Validation	31
2.3.3	Scores on Classification and Regression	32
2.4	Concept Drift	33
2.5	Machine Learning Explainability	35
2.5.1	Taxonomy	35

2.5.2	Explainability methods	35
3	Results	47
3.1	Predicting the Gender of Individuals with Tinnitus based on Daily Life Data of the TrackYourTinnitus mHealth Platform	48
3.1.1	Introduction	48
3.1.2	Results	53
3.1.3	Discussion	57
3.1.4	Materials and Methods	61
3.2	Prediction of Tinnitus Perception based on Daily Life mHealth Data using Country Origin and Season	69
3.2.1	Introduction	69
3.2.2	Materials and Methods	73
3.2.3	Results	79
3.2.4	Discussion	89
3.2.5	Data availability	95
3.3	Self-Assessment of Having COVID-19 with the Corona Check mHealth App	96
3.3.1	Introduction	96
3.3.2	Related Work	97
3.3.3	Technical Details	100
3.3.4	Results	105
3.3.5	Limitations	111
3.3.6	Discussion	111
3.4	How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare.	114
3.4.1	Introduction	115
3.4.2	Related Work	117
3.4.3	Explainability Methods	119
3.4.4	Materials and Methods	126
3.4.5	Results	130
3.4.6	Discussion	135
3.5	7 observational mHealth studies and 10 years of experience: Can ignoring groups in Machine Learning pipelines lead to overestimation of model performance? Analyses of group-wise validation as well as baseline and concept-drift considerations.	140
3.5.1	Introduction	141

3.5.2	Materials and Methods	145
3.5.3	Results	154
3.5.4	Discussion	159
3.5.5	Data and Code availability statement	161
3.5.6	Acknowledgements	162
4	Discussion	163
4.1	Recap of the research questions	163
4.1.1	What is the medical contribution?	168
4.1.2	What is the informatics contribution?	170
4.2	Overall interpretation of results	172
4.3	Limitations of this work	175
4.3.1	Model-related limitations	175
4.3.2	Data-related limitations	175
4.3.3	Domain-specific limitations	176
4.4	Future research	177
4.5	Conclusion	179
5	Appendix	181
5.1	Publication list	181
5.2	Affidavits	184
5.3	Curriculum Vitae	185
5.4	Contribution Statements	186
	References	191

Introduction

In recent years, the begun convergence of computer science and medicine has led to a transformative paradigm shift in healthcare delivery, leading to a research field known as **Digital Medicine**. At the same time, an Artificial Intelligence (AI) winter has been overcome by increasingly faster, cheaper, and more readily available computational power, which has led to an exponential increase in interest from the general public and academics in machine learning [1]. The advantages of merging computer science and medicine into digital medicine are apparent. Digital medicine can offer personalized healthcare through remote monitoring, data-driven insights, and patient empowerment, leading to early detection, reduced costs, and improved clinical decision-making. Its potential to bridge healthcare disparities, support research, and prioritize patient-centered outcomes underscores its transformative impact on healthcare delivery. Until now, however, many of these benefits have remained theoretical.

With Machine Learning (ML) as a subfield of Artificial Intelligence (AI), one has the tool to build high-dimensional early detection systems that can predict critical events regarding patient care, driven by data. However, this requires **data**, and the interplay of data generation, processing, interpretation, and prediction, especially in a clinical workflow, is not trivial. Fortunately, collecting data has become easier nowadays due to the high availability of mobile devices. There is the concept of **Mobile Crowdsensing**, a collaborative data collection approach that leverages the ubiquity of mobile devices to gather real-time information from a large and distributed group of individuals. If these individuals fill out assessments (synonym: questionnaires) about their current status, such as feelings, pain, thoughts, then these assessments are called **Ecological Momentary Assessments** (EMA). They add dimensions to the collected data, and if filled out over a longer period, and combined with sensor data from the mobile devices, such as the global positioning system (GPS), or microphone data, they form a high potential, longitudinal,

and multi-modal dataset.

ML can then help to utilize that data by finding new or confirm known associations between input and output, or predict future outputs based on current inputs. If one can show that an ML system can add value by meaningfully predict something, the system needs to be robust, confident, reliable, transparent, and trustful, among others. **Explainable Artificial Intelligence (XAI)** can help to reach these requirements by enabling humans to understand how a model makes certain predictions.

Digital Medicine, Mobile Crowdsensing, the challenges of using and analyzing mobile collected data with machine learning, and the explanation of these ML systems are issues addressed in this thesis. Chapter 1 provides a theoretical overview of the key concepts, formulates the research questions, and gives a brief introduction to the medical domains touched upon in the course of this thesis. Chapter 2 then establishes the theoretical foundations in machine learning and its evaluation, introduces a taxonomy for machine learning explainability with commonly used XAI methods, and explains the necessity of feedback loops with domain experts to collect multi-modal and longitudinal data. The thesis-contributing papers are subsequently listed in Chapter 3 and address the research question that are stated in the introduction. Chapter 4 summarizes the results of the research questions, discusses limitations, and provides suggestions for future research. Finally, a conclusion is given.

1.1 | Explanation of key concepts of this work

In this section, key concepts of this work are quickly introduced and defined. Some of them are defined citing related work, others are defined within this work here. This work embraces a total of six key concepts. To avoid confusion, note than in this work we use the terms *questionnaires* and *assessments* interchangeably.

Ecological Momentary Assessments In contrast to retrospective global self-reports, Ecological Momentary Assessments (EMA) are in real time and within the subjects natural environment [2]. The repeated sampling of a subject's natural environment aims to minimize recall bias and maximize ecological validity. It further allows for a long-term analysis of a subject's behavior change over time. Following the definition of Shiffman et. al., one can state that *EMA are methods using repeated collection of real-time data on subjects' behavior and experience in their natural environments* [2]. Key features of EMA are [3]:

- The ecological aspect of EMA means that the data is collected in real-world environments of the subject's lives. This allows for generalization of ecological validity.

- Moments are selected by random sampling at random points of time, or they are selected strategically based on events of interest, i.e., a relapse of mental illness.
- Assessments (i.e., questionnaires filled out using mobile devices) focus on the subject's *current* feelings, i.e., *What is your mood right now?* as proposed in the daily questionnaire within the TrackYourTinnitus project [4].
- Individuals (Depending on the study, this can be patients or app users) complete multiple assessments of time which provides an impression of a individuals' change of behavior on a longitudinal axis.

Early studies making use of the advantages of EMA include smoking relapse processes [5; 6], or more recent studies, the tracking of stress, tinnitus symptoms [7] or psychological health during the pandemic [8].

Mobile Health Nowadays, EMA are mainly collected using mobile devices. If, additionally, EMA are subject of medical studies, we are in the field of *Mobile Health* (mHealth). mHealth uses portable devices to create, store, retrieve, and transmit data for the purpose of improving quality of care [9]. mHealth apps mainly have the purpose of assisting, monitoring, informing, and educating, where the former two are implemented more often than the latter [10]. Diseases addressed by mHealth applications include diabetes, asthma, depression, hearing loss, low vision, osteoarthritis, anemia, and migraine [10]. During the pandemic, the coronavirus was the most trending topic within mHealth applications [11]. The combination of collecting EMAs using smartphones with wearable devices like smartwatches allows for a potential powerful creation of both multitudinal (across different sources, synonym: multi-modal) [12] and longitudinal (for a longer period) data sources [13]. Using different methods such as statistics, and machine learning, combined with the knowledge of subject matter experts (synonym: Domain experts), these data sources can then be applied to learn more about the disease and its individual-based branches. However, as [12] points out, most of these approaches still lack experience in deploying and maintaining these models.

Mobile Crowdsensing Mobile Crowdsensing (MCS) describes the collection of data with many mobile sensors from different mobile devices [14], often with contribution of many individuals that constitute a *crowd*. Two main differences from mHealth are the primer purpose of the collection: Firstly, mHealth is located within a health domain whereas MCS is not necessarily, and, secondly, the way the data is collected: Within the mHealth domain, users of mobile devices might create the data manually whereas in

MCS, data is collected automatically using sensors. MCS devices can sense, compute, and communicate, i.e., by sending aggregated data to a database. Within a personal sensing application, which can be seen as an overlap to the mHealth domain, data from a single individual or device is collected. In contrast, community sensing (synonym: participatory, or opportunistic sensing) has the aim to track large-scale use cases where multiple mobile devices are necessary, i.e., when determining the air pollution within a city. MCS applications can be limited by energy, i.e., the battery of a mobile device, and the computation capability, i.e., when data is aggregated on-edge before being sent to a server. Due to the large variety of mobile devices and operating systems, the quality of data or the way it is saved might also vary within a use case, even when the same sensor, i.e., GPS, is utilized [15].

Supervised Machine Learning Say one has an Input A which shall be mapped to an Output B . In Non-machine-learning systems, it requires Domain Expert knowledge to map A to B . One would need to bring the expert knowledge into the system which is why these systems are referred to as **expert systems** [16] or symbolic artificial intelligence [17]. On the contrary, when utilizing machine learning (ML), the system itself learns the A -to- B mapping if labels for the target are provided. In statistics, the target is also referred to as the outcome, or endogenous variable. Within the ML community, the target is also referred to as label. However, precisely spoken, a label is the value that the target can have, i.e., "female" and "male" for a classification task. As an example, a label could be an answer to a question within an assessment or the score of a PHQ depression questionnaire. If a ML use case is provided with labels, we refer to this use case as **supervised ML** [18]. Vice versa, we refer to unsupervised ML for unlabeled use cases [17].

ML algorithms can learn non-parametric A -to- B mappings. By *non-parametric* we mean that there is no assumption about the underlying distribution of the data or in other words, we do not make any assumptions about the connection of output B and input A . Because these underlying functions are non-parametric, this is referred to as a bias-free estimate of the target. Most of the use cases from this work has been addressed using non-parametric ML algorithms. Not every ML algorithm is non-parametric. Well known algorithms that are parametric would be Logistic Regression, Linear Regression, and Naive Bayes [17].

Machine Learning Explainability The ability of ML algorithms to learn complex A -to- B mappings make them opaque for those developing and applying them. At this point, Machine Learning Explainability, or Explainable Artificial Intelligence (XAI) comes

in. We define a XAI methods as follows: **XAI methods enable humans to understand why a model makes certain predictions** [1]. They can be either *local* (explain a certain prediction) or *global* (explain the whole modal behavior), and *model-specific* or *model-agnostic* [19; 20; 21; 22]. Model-specific XAI methods can be applied to specific kinds of algorithms only, whereas model-agnostic XAI methods are not limited to certain kinds of algorithms. To give examples, two widely adopted XAI methods are Shapley Value Explanations (SHAP) [23], mostly applied as a model-agnostic tool for tabular data, and Gradient-weighted Class Activation Mapping (GRAD-cam) [24], a post-hoc XAI method for image data that can be applied to neural networks to highlight pixels that have been relevant for the output of the model. XAI, its applications and limitations are explained in more detail in section 2.5 of this work and in the context of a literature review in section 3.4.

Domain Expert As the name suggests, domain experts are specialists in the domain the methodology is applied to. They are sometimes referred as *subject matter experts* [25]. Within the Cross Industry Standard Process for Data Mining (CRISP-DM), which is explained in more detail in subsection 2.3.1, the methodology expert and domain expert iterate multiple times to clarify the use case and data issues before further analysis can happen. When applying ML algorithms, domain experts often know causal relations between input A and output B which can be valuable starting points when developing an algorithm. Good and close communication between domain and methodology expert also favors progress in the use case [26]. When using XAI methods, domain experts can process specialist explanations, whereas the methodology expert works more with technical explanations to improve the algorithm.

1.2 | Problem statement and motivation

After having introduced the key terms of this work, we would like to motivate key challenges that are related to the key concepts that are described above. EMA in combination with MCS allows for multi-modal and longitudinal data collection, and the creation of large data sources. Without going into more detail about the term "large" in this context, we would like to discuss more about the inherit complexity that multiple data sources, once unified, have. This inherit complexity brings us to the first subsection of this chapter.

1.2.1 | What are specific challenges of data in the context of Ecological Momentary Assessments and Mobile Crowdsensing?

We identified 7 challenges that arise when working with EMA and MCS data. The list of these 7 is not exhaustive and they are partly linked to each other, and furthermore not necessarily all are EMA-specific. However, with the experience of published work, we found that they are likely to arise in the context of EMA and MCS data [27; 28; 29; 30]. These 7 challenges are **power users**, **no equidistant measurements**, **concept drift**, **missing values**, **measuring inaccuracy**, **user identity**, and **different user behaviour**.

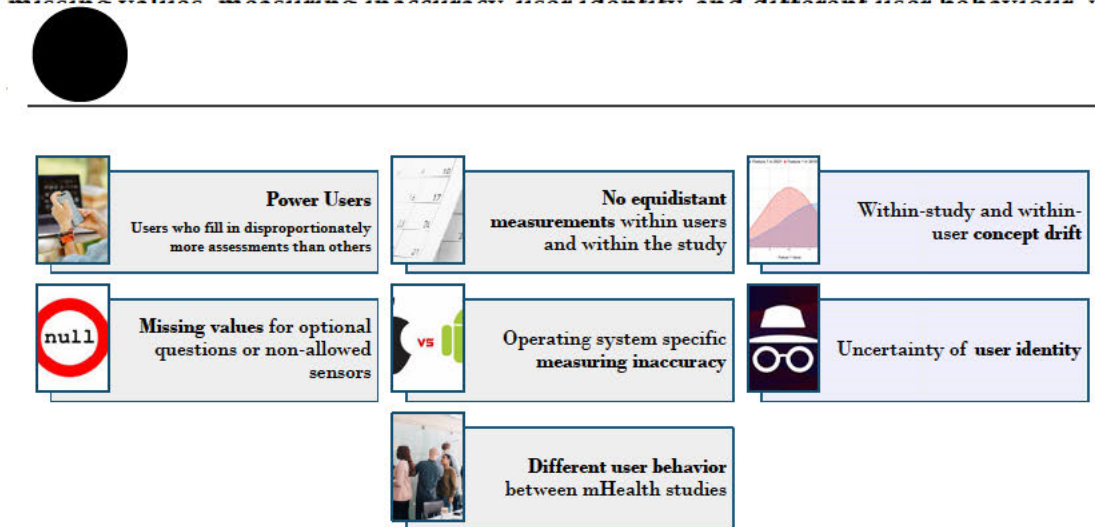


Figure 1.1: Seven mHealth challenges that are specific for longitudinal and multi-modal mHealth data. These challenges are going to be addressed in the following chapters.

Power Users The constant collection of EMA, especially in the users' daily environment, does only work for a small fraction of the study-involved users if there is not (monetary) incentive for the users to fill out these questionnaires. A very small fraction of the users however stick with the study for over 100, sometimes well beyond 1000 filled out assessments as one can see in Figure 1.2. These users are called **power users** [31]. There is no hard number of filled out assessments that need to be done before a user is considered a power user. It is rather substantially more assessments than most of all other users have. The challenge, now, arise from the skewed distribution and the bias induced into the dataset by these power users. If you train a model without paying attention to this skewed distribution (as an example, see figure 1.3), the model will learn to predict a power users behaviour rather than generalizing from the whole dataset which can be seen in more detail in section 3.2.

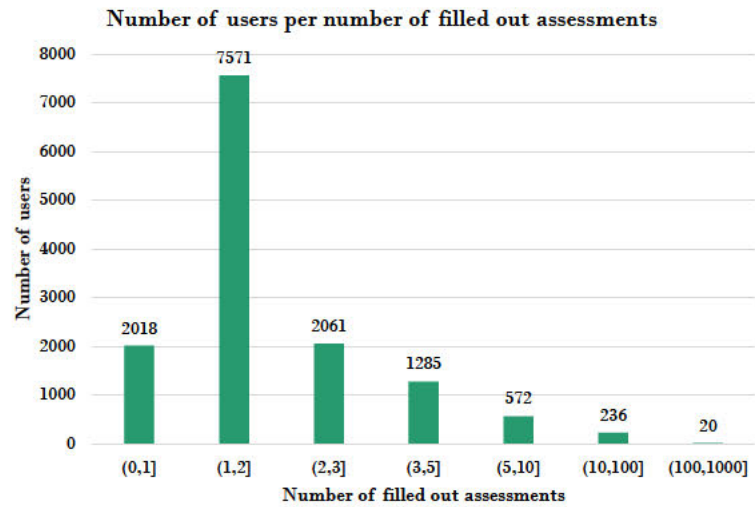


Figure 1.2: Distribution of number of filled out assessments. Most of the users fill out the questionnaire twice before dropping out of the study. We observed the behavior of early dropout in all studies involved in this work.

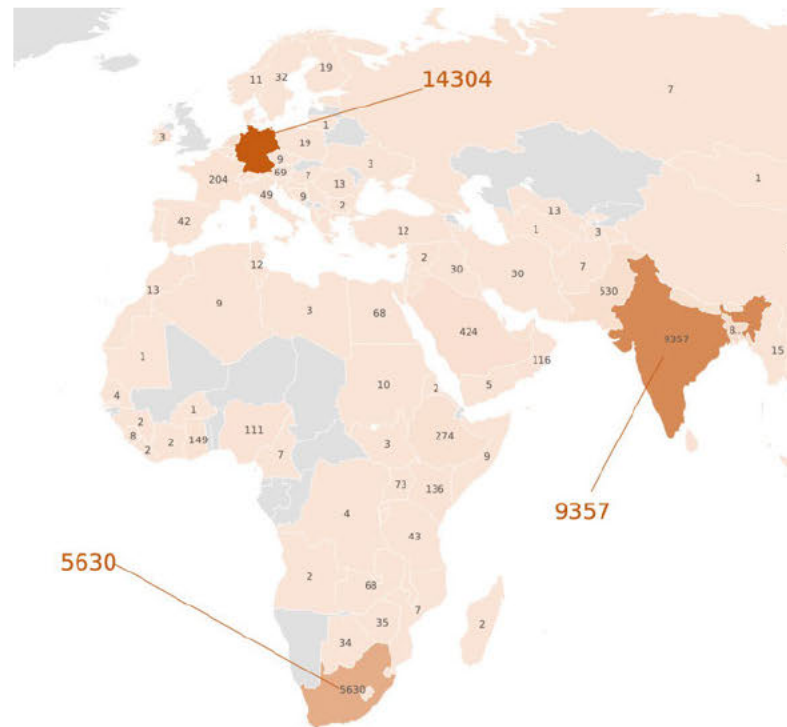


Figure 1.3: Example of skewed country-data distribution as by the Corona Check app. The numbers represented the number of users in this country, with Germany, India and South-Africa on most represented countries.

No equidistant measurements In most cases, EMA studies have a **baseline assessment** in which baseline statistics like age, sex, country, family status, education level, among others are asked to assess the generalizability of the users involved in that study. Then, these studies also have a **follow-up** assessments that are scheduled. As explained earlier, these schedules can be fix, i.e., daily at 8 a.m., or, within a specific period, random. Here, randomness however does mean within a specific period, i.e., at a random time between 8 a.m., and 8 p.m. This is the time a push notification will be displayed on a mobile device asking the user to fill out the EMA. If the user would immediately fill out the assessment, the assessment (synonym: measurement) can be considered *equidistant*. Voluntary and uncompensated users, however, do not usually fill out the assessments immediately, but sometimes only after several hours, days, or weeks, which can be seen in Figure 1.4. This brings up the next challenge: If a research question aims to collect assessment on a daily or weekly basis, this basis might be daily or weekly at the median of all users, but with a large standard deviation of the time gaps between two assessments of one user. More details about this are given in section 3.5.

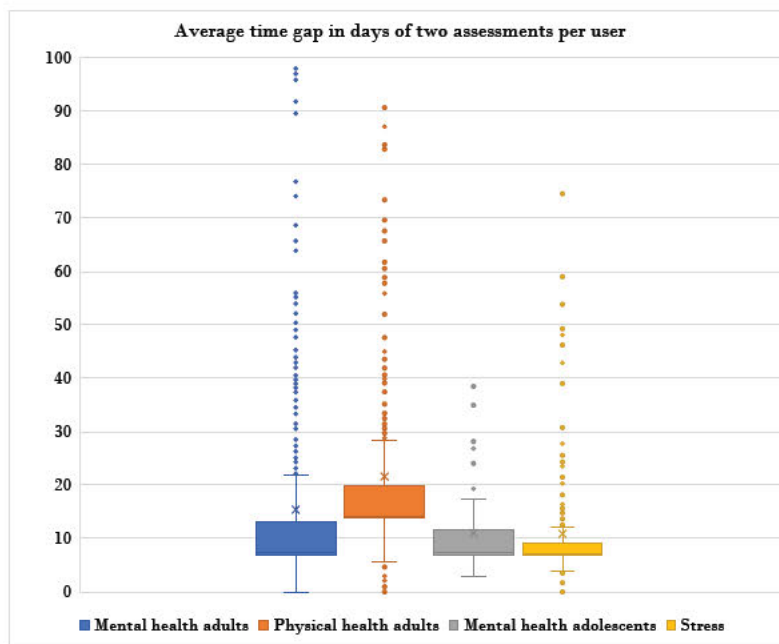


Figure 1.4: Boxplots of the average time gaps of two filled out assessments per user. The X indicates the mean, the horizontal line in the boxplot the median. The median value indicates the proposed time gap by study design (7 and 14 days). Because some users fill out the next assessment many days after due date, the mean values are right-shifted and add noise to the time dimension of the analyses. The four studies are introduced in more detail in section 2.1. For better readability, the limitation of the y-axis of this plot is set to 100 whereas the max value of the data is more than 300.

Concept Drift A deployed model is mostly trained with historical data. As described earlier, the model has learned the mapping of input A to output B . Within this paragraph, we refer to this mapping as *concept*, also sometimes referred to as posterior probability. This concept can then be described as a joint probability distribution with P denoted as the probability, input A and output B : $P(A, B) = P(B)P(A|B) = P(A)P(B|A)$. Technically spoken, concept drift happens if the probability of B given A changes such that $P_{t_1}(B|A) \neq P_{t_2}(B|A)$ with an older timestamp t_1 and a younger timestamp t_2 . Concept drift sometimes is referred to as concept shift or model drift [32]. Typically, real-world examples for such concepts are weather predictions or customer preferences [33]. Closely related to concept drift is data drift. In data drift however, the distribution of the input A changes over time but not necessarily the mapping from A to B . Given that A is a set of covariates, an example for data drift could be a covariate shift where the relationship of two features (synonym: covariates) changes within A . To give a real-world example, the unit measurement of a sensor could be changed after an update from inch to cm, or the sensor measuring accuracy decreases over time. Note that this does not change the mapping from A to B but the relationship within A .

Specifically, for the mHealth studies involved within this work, we potentially face both challenges, concept and data drift. We distinguish here between **within-study** and **within-user** concept drift. Since some power users stay in a study for several hundreds of assessments, they might change their mobile device in between which causes sensor measurements such as microphones to change, which states an example for within-user concept drift. Also, ecological circumstances might change over time. The mental health study of the Corona Health project, i.e., asks users about their feelings. Since the lockdown bylaws of the government changed within very short periods of time, this caused within-study concept drift.

Missing values Values can be missing in several sources. The most nearby might be non-filled out questions from assessments. If a question is not required to be filled out to finish filling out the whole questionnaire, users might skip this question. Another source are blocked GPS tracking or app tracking. Sometimes, i.e., research questions need this multi-modal data source to be correlated to the score of a patient health questionnaire. Regarding non-answered questions, there exist several methods how to treat these [34]. For GPS or app-usage-data however, these assessments must be excluded from the analysis which sometimes substantially decreases sample size.

Measuring inaccuracy As described earlier, measuring inaccuracy can happen if the sensor's quality decreases over time. However, there is also a problem with comparabil-

ity of measured values [35]. Different mobile device manufactures assemble different microphones or other sensors such that the exact same value will have different values once stored into the database. Even if the identical mobile device is used, one might be in a user's pocket while the other is in the user's hand which causes the same environment to be differently evaluated. Measuring inaccuracy also occurs when using wearable devices [36]

Different user behaviour Within the same project and study, users might answer, i.e., a question about their mood differently even if they felt the same. Although an admittedly subjective question, which is one reason for different filling-out behavior, another reason is the **anchor bias** [37]. The anchor bias is a cognitive bias in which irrelevant information is used as a fixed reference point (synonym: anchor) for future decisions.

User identity The true user identity is not confirmed when filling out the questionnaires. This is partly due to design reasons, such as in the Corona Check study [38], and partly due to inherent reasons. The Corona Check study has one question "Do you fill out this questionnaire for yourself or another person?". Within this assessment, there are further socio-demographic questions which are, as a pool, identity terminating such as sex, age, and nationality. Now, there are users in the dataset that filled out the questionnaire for themselves, but have different ages, sexes, and nationalities. The user identity thus is not a user identity anymore but rather a device identity. User-related analyses are, consequently, harder to address and one needs to make specific assumption when a device identity refers to a single individual. An inherent reason that obfuscates user identity is the fact that we do not know who is actually in possession of the mobile device while data is being recorded.

1.2.2 | Why is machine learning beneficial for evaluation?

Among the relatively younger machine learning approaches that can be used to address use cases and research questions, there are well studied statistical and mathematical methods that can be used to investigate. So, the question arise why one would prefer a ML approach compared to a simple heuristic or a rather complicated statistical method. Since the implementation and maintenance of an ML algorithm in production can be very elaborately, it is important to weigh whether the added value of the output of an ML algorithm exceeds the effort of implementation.

This depends on multiple factors. First, what is the type of input data. If the input data is **image** data, the effort with classical computer vision is also large, since features must be

derived manually from the images and programmed into an expert system. Here, modern deep learning methods with their end-to-end approach are advantageous. *End-to-End* in this context means that there are no requirements for feature engineering [39]. When the input data is tabular, simple heuristics may induce a bias in the *A-to-B* mapping, but they are robust, easier to understand, and require less maintenance. In some contexts, these heuristics are considered as **baseline models** [40] which is discussed in more detail in section 3.5. A ML approach can be considered promising, if its performance is strictly better than the baseline model, and if the costs of implementation and maintenance are lower than the benefits of the predicting algorithm.

Potential benefits of ML models In this paragraph, we would like to point out the potential advantages of ML techniques when evaluating use cases that have EMA data as input. In short, these points are:

1. **Complexity** | High dimensional and multi-modal data might be too complex for parametric models.
2. **Automation** | Potential value added in case of deployment of a model, e.g., for semi-automated clinical decision support systems.
3. **Impartiality** | Bias-free search for associations and correlations using non-parametric models.
4. **Novelty** | High demand from Domain Experts to address known problems with new methodologies like ML.

Complexity The preceded advantage of simple heuristics implies a high bias to the *A-to-B* mapping. The bias-variance-trade-off implies the balance of over- and underfitting the data. The model shall be complex enough to learn the *A-to-B*-mapping and yet simple enough to avoid fitting on noise or over-represented groups [41]. Simple heuristics, following the bias-variance-trade-off, thus have a high risk of being too simple to meet the complexity of the use case. When it comes to image data, simple heuristics within classic computer vision approaches further require a high expertise in feature engineering and standardization of processing image data.

This brings us to the second point, the potential value added of a deployed model, i.e., for clinical decision support systems. A major problem in clinics is the lack of staff time to provide care. Automation can relieve the workload of clinical staff by, for example, prioritizing patients according to their medical status, creating pre-diagnoses or classify x-ray images [42; 43].

Automation As described earlier, one induces a bias when using simple heuristics. A simple heuristic ("The diagnosis of this week equals the diagnosis of the next week") as well as some simpler ML models, such as exponential distributions, poisson distributions, normal distributions, the Weibull distribution, and linear regressions, are also referred to as **parametric models**. The mathematical formula of such parametric models is pre-defined, only the parameters are tuned during training. When using a parametric model, the ML engineer automatically must make an assumption about the underlying A -to- B mapping. However, when using **non-parametric** models, there is no need to make such an assumption. The task for the algorithm then is to find the underlying A -to- B mapping without having any parameters set before. Examples for non-parametric models are k-Nearest Neighbors [44], any kind of tree-based algorithm such as CART [45], C4.5 or Random Forests [46], as well as Support Vector Machines [47].

Impartiality The non-parametric models further allow for a bias-free search for potentially unknown associations and correlations from features within the set of input A to output B . Note that we did not use the word *causality*. Association is given, if a change of a variable x leads to a change of any property of y , i.e., the variance of y increases. Correlation, however, is given, if an increase of x leads to an increase or decrease in y . Because of spurious associations, causality can be confused with associations. As an example, one might drink 4 cups of coffee each day and has a decreased risk in developing skin cancer. There is an association between these two features. However, the true reason for the decreased probability of developing skin cancer might be that people who drink 4 cups of coffee per day work in the office many hours and thus are less exposed to the sun, which is a known risk factor for developing skin cancer [48]. Using ML algorithm and evaluating them using ML explainability and a hold-out test set, one can show that there is a correlation of input A and output B . However, this does not imply causality. However, deriving causality using ML is still under research investigation within the community [49; 50; 51].

Novelty Within the medical domain, many research questions are unsolved to date. At the same time, ML methodology is still a reasonable young research fields with exponentially growing interest within recent years. This leads to a high demand from domain experts to try out new methods on unsolved research questions, i.e., within the TrackYourTinnitus project, which will be explained in more detail in section 2.1. For example, using ML explainability, the ML algorithm might detect a unknown association of input A and output B might then cause a change of the focus of future work [52].

1.2.3 | Why is machine learning explainability needed?

After having motivated the key concepts of this work and the potential benefits that come with the usage of ML, we would like to motivate the need of machine learning explainability within the medical domain in general and, using tabular data and EMA within a supervised ML use case, in particular. In short, factors that motivate the use of ML explainability are as follows:

1. Reconciliation of the ML findings
 - with the knowledge of the domain experts [53],
 - against the models themselves (cross-validation) [54], and
 - against simple (non-ML) heuristics [55].
2. Model debugging and model performance improvement [56].
3. Black box character for big ensemble methods: building trust for the system [57].
4. Compliance with regulatory restrictions [58].
5. Confirmation of the finding of new correlations through traceability of the predictive behavior.

Reconciliation of the ML findings Given that the ML model finds an association between input A and output B , which is generally the case if model performs significantly better than a baseline model or heuristic, one can affirm the found association with different stakeholders. One of these stakeholders is the domain expert, who can affirm a given output of the model if feature importance is provided by a explainability method. To give a concrete example, a feature importance method may detect the number of pregnancies of a patient for a diabetes-detection algorithm [59]. The domain expert can then certify that the number of pregnancies is indeed a contributing factor for developing diabetes which in turn means that the model has learned something meaningful.

Using **cross-validation**, which is explained in more detail in section 2.3.2, one will get different performance scores of an algorithm within each validation fold. If the performance scores' variation is above a certain threshold, this is an indication of overfitting [60]. One can also apply global machine learning explainability methods to derive what the model has learned in each training fold. Analogous to the performance scores in the validation folds, a high variance in the feature importance ranks can be used to infer overfitting of the model.

A third option is to validate the rather complex ML models against simple heuristics

using ML explainability. If the ML explainability methods extends the simple heuristic, this can also be interpreted as a sign that the model has learned something meaningful. To give a concrete example, a simple heuristic for a weather forecast could be: "The weather of tomorrow equals the weather of today". If a machine learning explainability method states "today's weather" and some humidity and airflow features as the most important contributing features, it has extended the complexity of the heuristic to improve model performance and further confirms the heuristic which is based on domain expertise.

Model debugging and model performance improvement A method to systematically improve model performance is to carry out error analysis. A neural network may be trained to classify dog breeds. Carrying out error analysis, one may find that the network confuses huskies with wolfs. If now additionally a ML explainability method such as Grad-CAM [24] is applied, there might be highlighted snow in a saliency map which means that the model associates both wolfs and huskies with white pixels. This finding now implies to collect more images of wolfs and huskies in a non-white setting to help the model generalizing further. With the error analysis alone, one might have understood that there *is* a problem with wolfs and huskies, but not *why*.

Building trust for the system There are scenarios where ML explainability is not required, for instance, in a logistics center, where packages are classified according to the place of shipment. If the model performs well in a deployed scenario and might misclassify one in one thousand packages in average, the Chief Executive Officer would not care *why* as the benefits of the model exceeds costs caused by rare errors by far. Within the medical domain however, stakeholders, which are listed in more detail in section 2.5, have a great interest of why a model makes certain predictions for various reasons. Addressing the correct depth and technical level of explainability to the stakeholder of interest can increase trust of the ML system and thus increase the likelihood of software integration [61; 62; 63]. One keyword to mention here is the **human-in-the-loop** approach. Through the black box character of most clinical decision support systems that are ML-based, clinicians need to provide feedback for the ML system and correct false outcomes or confirm predictions not only to improve future predictions, but also to increase the trust in the system.

Compliance with regulatory restrictions So far, regulations for AI systems are still emerging worldwide, such as the *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government* (US Executive Order 13960) [64], the *White House Blueprint for an AI Bill of Rights* [65], and the *AI principles*, issued by the Organization for Economic

Cooperation and Development (OECD) [66]. The regulations are partly being formulated with the help of expert groups and in feedback loops with the scientific community [67]. Since the definitions of AI and ML as well as the areas of application are very heterogeneous and also change with great dynamism, the conflicting goals of regulations are to provide concrete specifications without limiting the potential of the systems and, above all, protecting the individuals. Among other key aspects, all these regulations have in common that they address transparency and explainability.

Regarding regulation within the medical domain, for the United States of America (USA) the competent authority is the Food and Drug Administration (FDA), for the European Union (EU) it is the EU commission. Both authorities refer to AI in a broader context as *software as a medical device* [58; 68]. So, what is **software as a medical device**? The FDA defines this as *"software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device"* [69]. The EU commission uses a very similar term **medical device software** (MDSW) and defines it as *"software that is intended to be used, alone or in combination, for a purpose as specified in the definition of a "medical device" in the Medical Devices Regulation (MDR) or In Vitro Diagnostic Medical Devices Regulation (IVDR)". A medical device is then defined as "(...) any instrument, apparatus, appliance, software (...) intended by the manufacturer to be used (...) for human beings for one or more (...) specific medical purposes (...)"*. Differences of older version of the USA and EU regulations have been summarized and discussed [70; 71] about 10 years ago. With recent dynamically growing interest of a broader public, as well as new regulations to be published in 2024 [68], one expects more comparative work on this in future related work since this is out of scope of this thesis. The fact that all these drafted regulations and recommendations mention explainability highlights the importance of XAI methods.

Confirmation of the findings This paragraph has an overlap with the reconciliation of the ML findings. In a feedback loop with an algorithm-applying user, the domain expert can confirm a prediction of an algorithm based on his or her own knowledge with respect to the current case. This not only builds trust for the deployed system but has the potential to increase domain knowledge if a trusted system correctly predicts an outcome with an explanation (detail) that the domain expert did not expect.

We now have seen three major questions that motivate the research direction of this thesis: The EMA and MCS specific challenges, the benefit of using machine learning when working with this kind of data, and the necessity of XAI methods to explain and understand model predictions. After this introduction to the methodological topics of the thesis, we give an overview of the domains from medicine that are touched upon.

1.3 | Introduction to thesis-related domains

This section gives an introduction into the two domains of data that this thesis works with: Tinnitus, and the coronavirus. Sections 3.1, 3.2, and 3.5 work with a multi-modal and longitudinal dataset that contains mHealth tinnitus data from the TrackYourTinnitus research project which is introduced in section 2.1. The tinnitus data was also used to answer Main RQ 1 (*How can machine learning help confirming or broaden domain knowledge within mHealth data?*), and Main RQ 2 (*How can one reach explainability in the presence of mHealth data when using Machine Learning?*) Sections 3.3 and 3.5 work with data that was collected using the Corona Check mHealth app. The app was launched in an early stage of the COVID-19 pandemic in 2020 to help overburdened coronavirus hotlines and track the course of the pandemic using MCS and EMA. The dataset of the Corona Check app was then used to answer Main RQ 3 (*Which guidelines can be beneficial for the use of ML within the mHealth domain ?*).

1.3.1 | Tinnitus

Tinnitus, often referred to as "ringing in the ears", is a widespread auditory sensation characterized by the perception of sounds in the absence of external acoustic stimuli. It is estimated to affect approximately 10-15 % of the world's population [72]. While tinnitus is most described as ringing, it can also manifest as buzzing, hissing, clicking, or other phantom sounds. Tinnitus is a complex and multifaceted condition that can significantly affect a person's quality of life, leading to stress, sleep disturbances, and difficulty concentrating. Furthermore, this clinical picture exhibits a high degree of heterogeneity with many, sometimes unknown, influencing factors. The underlying mechanisms of tinnitus involve intricate interactions within the auditory system and the central nervous system [73]. However, tinnitus perception is also influenced by central auditory processing, neuronal plasticity, and emotional factors [74]. Furthermore, gender-related psychological and socio-cultural factors might contribute to variations in tinnitus reporting. This hypothesis is further investigated in section 3.1. Also, seasonal tinnitus was reported among tinnitus patients. That is, does tinnitus vary within a year's course based on temperature or season? This question is asked in [75] and further investigated in section 3.2.

1.3.2 | Severe Acute Respiratory Syndrome Coronavirus 2

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, short: Coronavirus), the causative agent of coronavirus disease 2019 (also known as COVID-19), emerged in late 2019 and rapidly escalated into a global pandemic, challenging public health systems, healthcare infrastructures, and societies worldwide. Characterized by a diverse range of symptoms, COVID-19 has exhibited an array of clinical presentations, varying from mild or asymptomatic cases to severe respiratory distress and multi-organ dysfunction. The coronavirus presents a wide range of symptoms that can manifest in diverse combinations and severity levels. While the most reported symptoms include fever, dry cough, and fatigue, other clinical manifestations have been identified. These symptoms encompass, among others: Shortness of breath, chest pain, persistent cough, diarrhea, loss of taste and smell, headaches, and muscle pain [76]. In section 3.3, we investigate, among others, whether the distribution of the symptoms differ between countries.

In this introduction, three essential things were shown. **First**, we have defined the six key concepts of this thesis (Ecological Momentary Assessments, Mobile Health, Mobile Crowdsensing, Supervised Machine Learning, Machine Learning Explainability, and Domain Expert) , distinguished them from homonyms and synonyms, and illustrated them with examples.

Second, we delineated the problem field and put it into context. Seven challenges that are not limited to EMA applications but may arise when working within this field were introduced and explained. We further answered the question why ML is beneficial for evaluation of data that is introduced in section 2.1 (potential benefits, complexity, automation, impartiality, and novelty). Finally, we explained the need of XAI within the mHealth domain. Analogously to the benefits of ML, these aspects (reconciliation of findings, model debugging, trust, compliance, and confirmation) do not necessarily apply in the mHealth domain only.

And **third**, we gave a short introduction to the thesis related domains, tinnitus, and the coronavirus, and how they related to the three main research questions. Within the tinnitus domain, this thesis investigates the heterogeneity picture of the tinnitus syndrome by investigating into gender- and season related differences. Regarding the coronavirus, we examine, among other things, the different distributions of reported symptoms among mHealth Corona Check app users.

Materials & Methods

In this chapter, we give an overview of the datasets and methods that are used in the papers that contribute to this cumulative thesis. The chapter is subdivided into three sections. Section 2.1 describes and summarizes the EMA datasets that are involved in all papers for direct or meta analyses. The section also briefly discusses how the technical pipeline runs in the background (subsection 2.1.1) to communicate with both Domain Experts and Back-end Developers to minimize information leakage regarding Domain Expert implementation requests and technical feasibility. Once the data is stored in a database, it can then be exported and visualized on a dashboard which will be explained in more detail in subsection 2.1.2. Section 2.2 introduces the three major subfields or machine learning: Supervised, unsupervised and reinforcement learning. As this thesis operates within the supervised ML field only, this field is explained in more detail. Also, major issues such as the bias-variance trade off or overfitting, and concepts like ML pipelines are explained. As this thesis uses random forests for prediction, tree-based ensemble methods are also explained. The chapter closes with a section about XAI (section 2.5). Here, a taxonomy for local and global explainability methods is given, as well as a list of common explainability methods with explanations, applications, advantages, and disadvantages for each method.

2.1 | EMA mHealth datasets

This thesis incorporates data from four **research projects**. Each research project has its own mobile app, developed on both iOS and Android. The Corona Health (CH) research project, for example, further includes several **studies**, and each study can have several questionnaires which we refer to as **assessments** in this work.

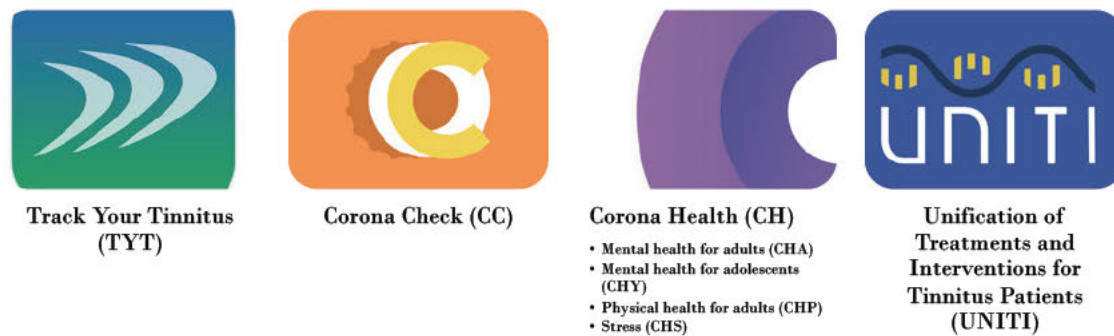


Figure 2.1: Research projects that use EMA which are involved in this dissertation. Track Your Tinnitus (TYT), Corona Check (CC), the Corona Health (CH), and Unification of Treatments and Interventions for Tinnitus Patients (UNITI). The images above the project names are the logos of these projects.

Dataset	No. of users	No. of assessments	First assessment from	Dataset span	Ø Age (Std)	Ratio M/F	% male GER users
TYT	3303	110983	2013-07-18	9.20	45.0 (14.4)	67/33/00	n. A.
CC	13763	89659	2020-04-08	2.48	32.7 (18.0)	59/39/01	36
CH	1474	11081	2020-07-21	2.19	41.2 (13.9)	54/45/01	98
CHA	1474	11081	2020-07-21	2.19	41.2 (13.9)	54/45/01	98
CHY	111	630	2020-08-08	2.14	15.2 (1.6)	51/47/01	n. A.
CHS	374	3845	2020-12-19	1.78	40.7 (13.9)	65/34/01	98
UNITI	763	32443	2021-04-13	1.46	53.0 (12.7)	57/43/00	54
Sum	20741	254302	n. A.	21.42	Ø 36.48	Ø 60/40/00	Ø 39

Figure 2.2: Baseline statistics of all 7 studies that are included in this thesis. More than 20,000 users filled out more than 250,000 assessments, with the earliest assessments starting in 2013. The weighted average age of a user is 36 years with a weighted, average ratio of 60-40 from male to female users. Users located in Germany make up 39 % of all users. Ø means number-of-users-weighted average.

In the next paragraphs, we would like to introduce the questionnaires of the studies in more detail. The focus of this introduction, however, is more on a meta-level than on a domain-level, since the different studies have different domain backgrounds and the focus of this work is rather on the evaluation side of these datasets than on the domain background.

TrackYourTinnitus (TYT) Started in July 2013, this study collects data for almost 10 years to date and is the most longitudinal study in this thesis. It contains four assessments of which three have been included in evaluation of this thesis. The three included are at first, a baseline assessment *Tinnitus Sample Case History Questionnaire* (TSCHQ), which asks for demographic data and tinnitus history of patients as well as the worst symptom associated with tinnitus. Second, a follow-up questionnaire (database-internally referred to as "standardanswers") on a daily basis that collects EMA from the users, asking for

their current mood and current tinnitus perception. Before being able to fill out the daily questionnaire, users must fill out the TSCHQ. Third, the *Worst Symptom* questionnaire which asks the user to report the worst symptom. This worst symptom is then asked for in the follow-up daily questionnaire "Do you perceive your worst symptom right now?". During the evaluation of this study, we had to be careful caused by unconventional data encoding to avoid creating outliers in the data. For example, missing values in the question about date of birth were coded as '???.??.????'. In addition, there were no plausibility checks for age information, so that some users were well over 100 years old, which is rather implausible when filling out an electronic questionnaire via smartphone. The time gap of two assessments from the same users is supposed to be 24 hours, on average.

Unification of Treatments and Interventions for Tinnitus Patients (UNITI) This study is also related to TYT because the same scientists were involved in its design, such as Prof. Rüdiger Pryss and Prof. Winfried Schlee. The questionnaire involved in this thesis from the UNITI project contains over 32,000 assessments filled out by 763 users. The questions within this assessment are highly correlated with the daily questionnaire from TYT, i.e., UNITI asks *How loud is your tinnitus at the moment?* and TYT asks *How loud is your tinnitus right now?*. A phenomenon already mentioned in the introduction is the large standard deviations of the users when filling out the assessments, the procrastination of answering. On average, users complete questionnaires every day, but the standard deviation is 4.6 days, which means that a good third of all questionnaires completed by a user are completed after 4.6 days. We will see in the main section that other studies have even larger standard deviations.

Corona Health (CH) The previous two projects mainly addressed the tinnitus disease whereas the next two projects address the COVID-19 pandemic and its psychological and physiological consequences caused by lock-downs and isolation, as well as the fear of a COVID-19 infection. Within this thesis, we included four studies from the CH project that are all hosted within the CH app. Three of these studies are psychological (mental health for adults, mental health for adolescents, stress), one is physiological (physical health for adults). 2,912 users filled out a total of 21,217 assessments. Each study consists of two questionnaires, a baseline and a follow-up questionnaire with time gaps between two follow-up questionnaires of one or two weeks. Exponential dropout rates as well as large standard deviations in the perceived time gaps between two assessments of one users make this data challenging to evaluate. The app design and the challenges of the app development are explained by Vogel et. al. [77]. The *Mental health for adults*

study was translated into 7 languages to enable as many people as possible to fill in the questionnaires in their native language. Since the network of scientists and thus the level of awareness of the study was greatest in Germany, also due to advertising and involvement of the Robert Koch Institute, the studies were mainly downloaded and completed in Germany. More details of the studies of the CH project are given section 3.5, app screenshots are given in Figure 2.3.

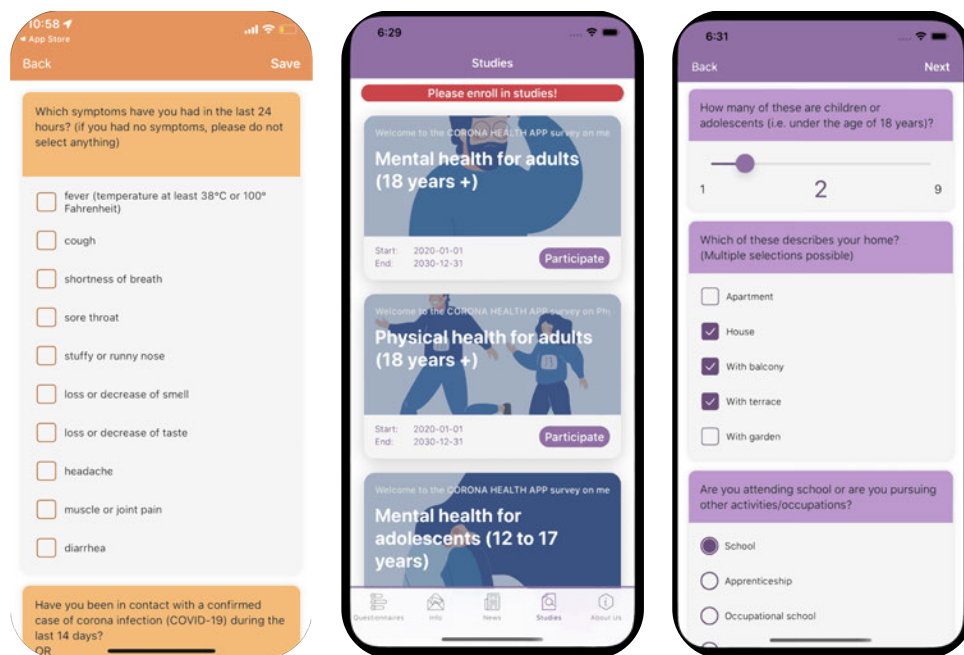


Figure 2.3: Three screenshots of the Corona-based EMA projects. The screenshot in orange colors is from the Corona Check project. Within this assessment, users can report symptoms which are used to provide recommendation for action such as self-isolation or the consultation of a physician. The screenshots in purple belong to the Corona Health project. The middle picture shows three of the four studies and gives users the option to enroll in one or more of these. The picture on the right is a screenshot of the *Mental health for adolescents* study. The apps were developed by Vogel et. al. [77].

Corona Check (CH) The goal of the Corona Check app is partly in its name, namely an initial heuristic check of potential Corona infection based on - at the time - known symptoms that correlated with infection of the virus which rules can be seen in Figure 2.4. The app could be downloaded from common app stores without further ado, and without creating a profile, the Corona Check questionnaire could be completed. At the time of the User-Assessment paper (see section 3.5) there were approximately 13700 users with a total of over 89,000 questionnaires.

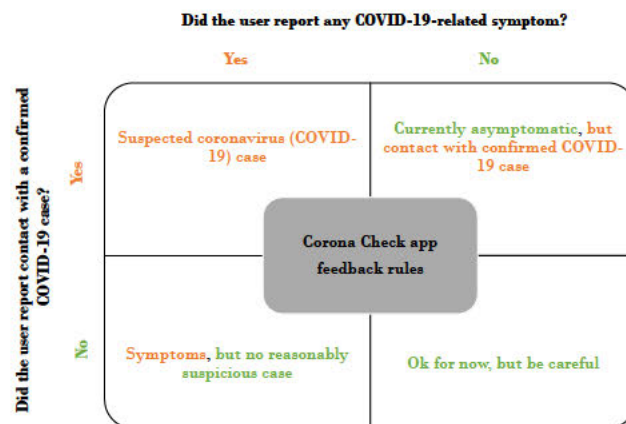


Figure 2.4: Corona Check evaluation matrix based on two questions: Did the user report any known COVID-19-related symptom and did the user report contact with a known COVID-19 case? Based on these answers, a user was given feedback and tips on how to proceed further.

The advantage of easy access to the app without access restrictions and direct filling of the questionnaire in combination with the question "Do you fill out the questionnaire for yourself or for another person?" drastically complicates the evaluation with regard to baseline characteristics and user identity. In contrast to the other projects CH, TYT and UNITI, which had baseline questionnaires to collect demographic data, the CC app did not. As a result, we later had to make assumptions that a questionnaire belonged to the same person. These assumptions are described in more detail in section 3.5.

2.1.1 | Excel-Loop

The creation of an assessment is an iterative process that involves several experts from different fields. The fact that different domain experts are part of this iterative steps brings up communication issues. Originally, domain experts sent us Microsoft (MS) Words documents or PDF files that included the questionnaires, ready for print-out. The backend however, designed by Johannes Schobel in his doctoral thesis [78], requires nested Java Script Object Notation (JSON) files which most physicians have never seen before. The task at hand was to structure the unstructured MS Word document then. We then had the idea to use an MS Excel file for this which is then converted into an Application-Programming-Interface (API) readable format. This process is visualized in figure 2.5. A section about the excel-loop is published in a technical paper that addresses the Corona Health study [8].

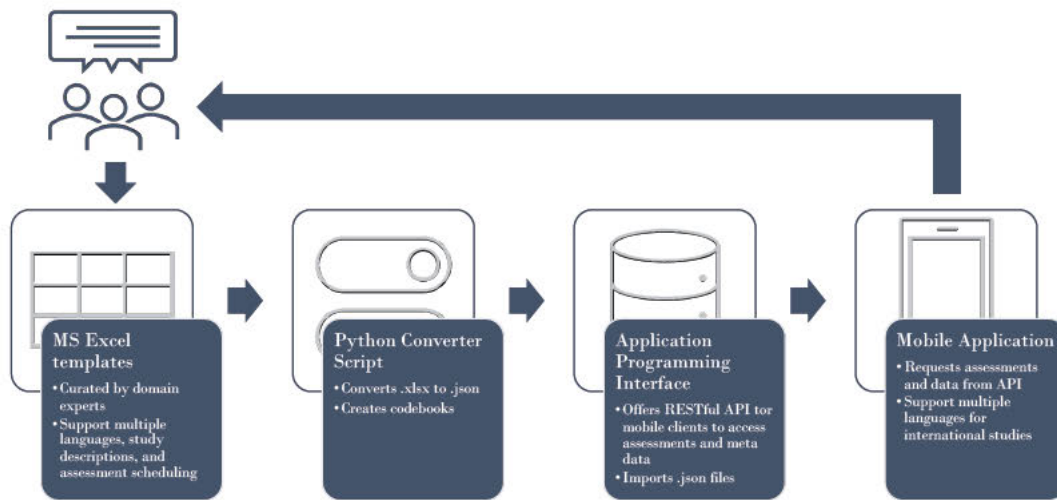


Figure 2.5: Iterative process of developing a study using mobile devices and excel templates that allow structured communication between study designers and programmers. Automation scripts speed up the process and communication between stakeholders as well as minimize information loss between groups. This figure is based on my own figure from [8]

Advantages of this half-automated pipeline are:

- Communication with domain experts using structured documents
- Quality assurance of questionnaires using unit tests, such as using the same key-value pairs of answers for multi-languages questionnaires
- Acceleration of the iterative questionnaire creation process
- The Excel sheet itself can be later used as a codebook to evaluate the questions
- Excel is one of the most used data formats to date between different research areas and industrial branches

I also provided a video tutorial on [YouTube](https://www.youtube.com/watch?v=wa-Fd4sjWg4)¹ that introduces the Excel sheets to new domain experts.

¹<https://www.youtube.com/watch?v=wa-Fd4sjWg4>

2.1.2 | The dashboard

The [dashboard](#)² is a follow-up project of the Excel Loop and has the goal to inform the domain experts of a study about key figures of the incoming assessment on a daily basis. In addition, graphics with various groupings and aggregations are given for core questions from the studies. For example, the current mood, grouped by age or gender or severity of illness. The dashboard project is presented in this project video on [YouTube](#)³. A screenshot of the dashboard is given in Figure 2.6.

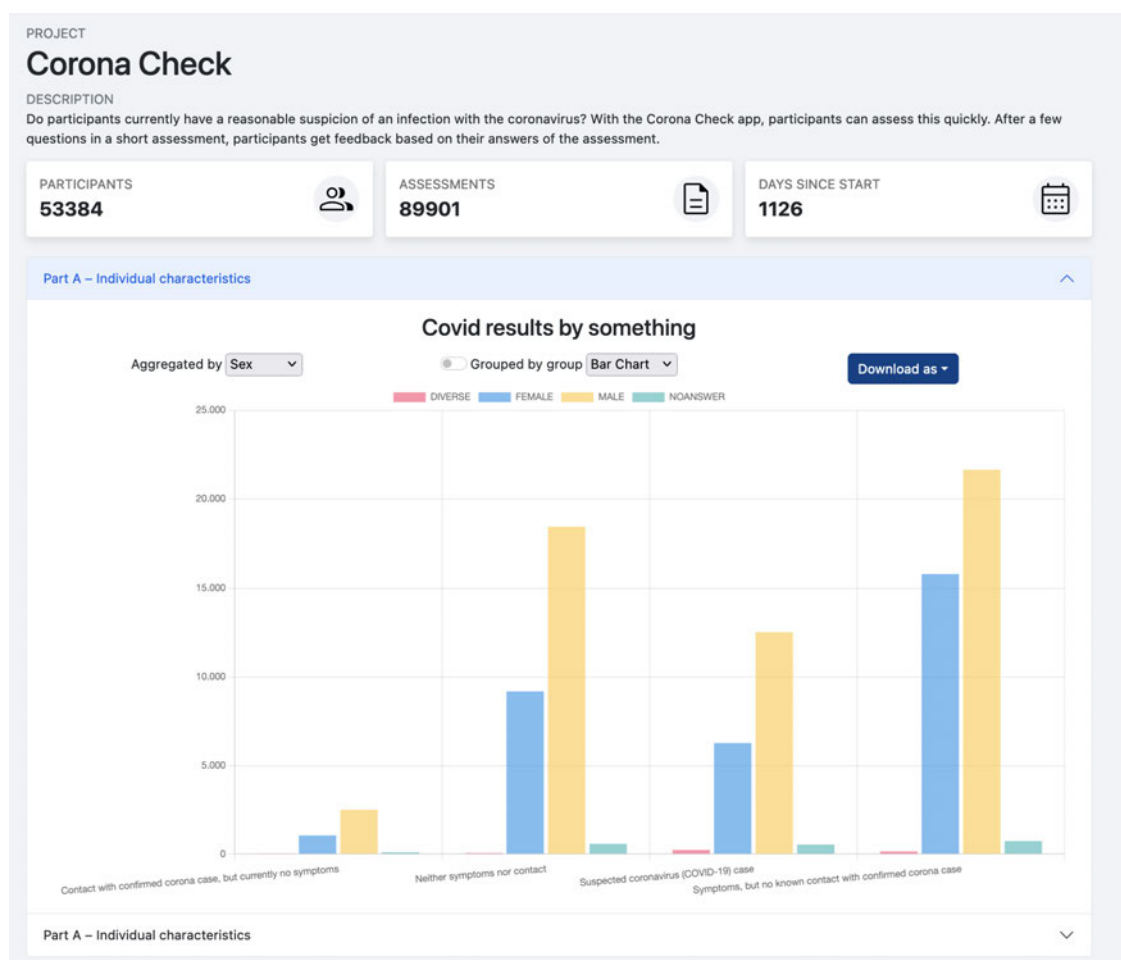


Figure 2.6: Screenshot of the dashboard project with data aggregation of the Corona Health project. The data can be aggregated by sex, age, and country. If the figure fits the users' needs, it can be downloaded as a PDF or PNG file.

²<https://mhealth-dashboard.de/>

³<https://www.youtube.com/watch?v=Dj7sU2vhe1c>

In this section, we introduced the data used in this thesis on a meta-level, gave facts and statistics about it, and explained the studies and mobile app integration for it. We also introduced the Excel loop that reduces the communication gap between programmers and domain experts of the medical and psychological apps by turning unstructured data into structured data and automating the conversion of the Domain Experts' information and requests regarding questionnaire design. Last but not least, we presented the dashboard that aggregates the study data generated by the Excel loop and redisplay it in graphs to keep the Domain Experts informed about the study progress on a daily basis.

2.2 | Supervised Machine Learning

Although this thesis uses Machine Learning in almost every contributing paper, it is not about Machine Learning. Nevertheless, in this section we want to give a brief introduction tree-based algorithms. The topic of machine learning has grown exponentially since the beginning of my thesis, not least due to the release of newer version of large language models like *GPT-3.5* and *GPT-4* by the company *OpenAI* [79] with early work starting in the late 1950s from Samuel [80]. Among supervised Machine Learning (ML), there exist two other types, unsupervised ML and Reinforcement Learning as shown in Figure 2.7. The most popular by far, however, are supervised use cases where model is given an input A and an output B as shown in section 3.4. Within supervised ML, the model has the task to learn the mapping from A to B . This is a substantial difference to classic programming, where an input A and a set of rules is given to determine B . Within the ML jargon, the output B is referred to as *targets* or *labels*. For unsupervised ML use cases, these labels are not given and the model task is either clustering analyses or dimensionality reduction. For this work and with the current state of the art, we would like to define Machine Learning as follows: **Machine Learning (ML) learns from an input A to an output B .** Within supervised ML, we differentiate between *classification* and *regression* tasks. Classifications are outputs on a concrete scale, i.e., pneumonia vs. not-pneumonia would be a binary classification task. The prediction of blood pressure, however, is on a continuous scale and therefore a regression task. There are algorithms that can only classify or regress, depending on how they are structured. One of the standard books in AI give a deeper introduction of these concepts [17].

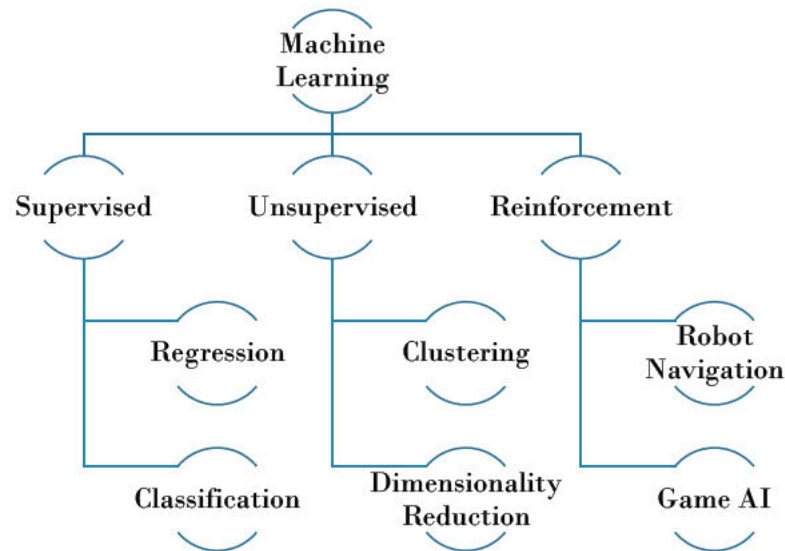


Figure 2.7: Subfields of Machine Learning. This thesis focuses on supervised ML with regression and classification tasks.

2.2.1 | Bias-variance tradeoff and the problem with overfitting

One of the big challenges for ML tasks is the **Bias-Variance tradeoff** [41; 81; 82]. The model shall be *rich enough to express underlying structure in data and simple enough to avoid fitting spurious patterns* [41]. In other words, the balance of over- and underfitting the data. If one overfits the training data, the model behaviour on deployment becomes unpredictable and performance will drop. If the model underfits the data, it is not complex enough to express the inherently learned A-to-B mapping. Section 3.5 discusses different cross-validation techniques (section 2.3.2) in order to avoid overfitting. The word *bias* here is not to be confused with bias in data selection. In our ML context, bias describes the assumptions that are made about the inherently A-to-B-mapping. For example, if one chooses a linear regression model, this induces a strong bias about the data distribution, namely mapping is linear. The goal of machine learning is to find an optimal balance between bias and variance to achieve good generalization performance. This can be achieved through techniques such as regularization [83; 84; 85], cross-validation [86; 87], and model selection [86]. Regularization methods such as L1 and L2 regularization can reduce variance by adding a penalty term to the model's objective function, discouraging overly complex solutions. Cross-validation allows the performance of a model to be evaluated on multiple subsets of the data to assess its generalization ability. Model selection involves choosing the appropriate level of

complexity for the problem at hand, often by comparing the performance of different models on validation data.

2.2.2 | Tree-based ML methods

Tree-based machine learning methods are powerful and widely used techniques for both classification and regression tasks. These methods build decision trees or ensembles of decision trees to make predictions based on input features. Quinlan introduces the concept of decision trees and outlines the ID3 algorithm for inducing decision trees from labeled training data [88]. He discusses attribute selection criteria and pruning techniques, such as Gini impurity, Entropy and Information Gain. These concepts measure the disorder of the data based on the target and try to maximize the order such that the target distribution has minimum entropy. Tree-based methods are widely adopted because they are interpretable and non-parametric, meaning one has not to make an assumption about an underlying distribution. Disadvantages of decision trees have been partly addressed by Breiman in his paper *Random Forests* [46]. He creates an ensemble of multiple decision trees, each tree trained with a slightly different subsample of the training data using a technique called bootstrapping. Because decision trees are greedy, each tree is built slightly different and thus creates different predictions for the same given input. Another difference of random forests compared to decision trees is the random feature selection. That is, not only a subset of the training data is given to each tree, but also a subset of the available features. Random forests are much more robust than decision trees because of the variance of the trees within a forest. The bootstrapping process of the random forests has been optimized in a follow up algorithm called gradient boosting machine (GBM) [89]. Random forests build each tree independently without considering the errors of previously built trees. In contrast, GBMs iteratively construct trees by focusing on the residuals or gradients of the loss function, aiming to reduce the overall prediction error at each iteration. GBMs further use all training data but weighted on the errors the previous tree made. Random forests, in contrast, always use a subset of the training data to construct a new tree. Although GBMs generally achieve a higher performance than random forests, random forests still remain popular because of their interpretability and their insensitivity to outliers. GBMs however, can be sensitive to outliers since they focus on misclassified data points during training.

2.3 | Machine Learning Pipelines

In this section, we would like to discuss ML pipelines in the extended and narrow sense. An extended ML pipeline is shown in Figure 2.8 and serves as an orientation for the use and deployment of ML algorithms; in the narrower sense, it serves mainly to ensure traceability and reproducibility of research using ML, which is explained in more detail in section 3.4.

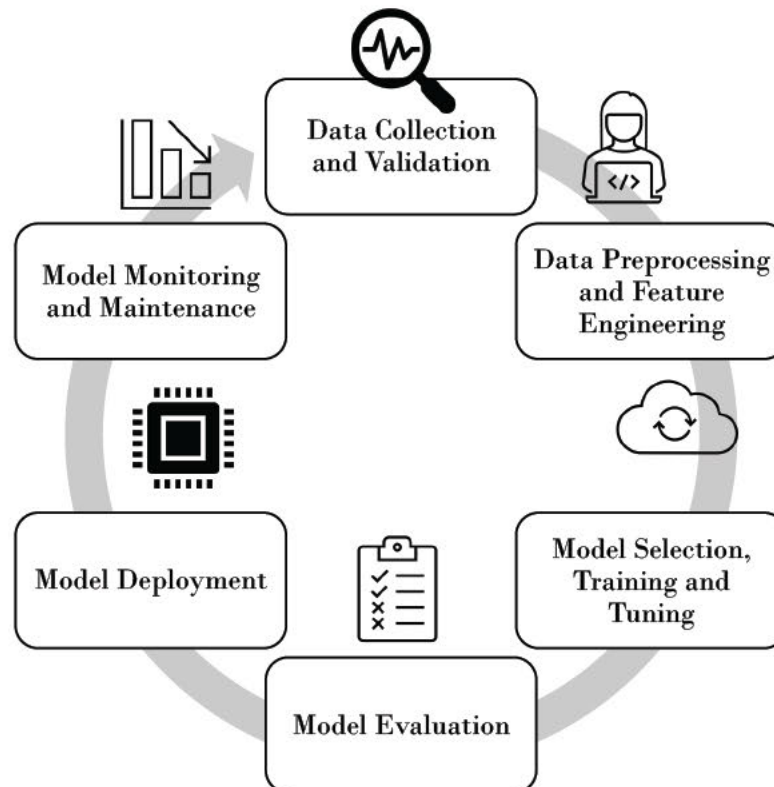


Figure 2.8: Exemplary image of a ML pipeline with 6 elements data collection and validation, data preprocessing and feature engineering, model selection, training and tuning, evaluation, deployment, and monitoring. Due to concept and data drift (section 2.4), continuous integration and monitoring is important to keep the model up to date. A whole machine learning pipeline includes specialists from various fields like software, security, governance, machine learning, data science, subject matter experts, and users.

ML pipelines in an extended sense include the steps that are necessary within a ML project to bring the model in production so it can generate output in a real-world scenario and add value. The whole life cycle of deploying a ML algorithm is a complex process that involves experts from different areas like software engineering, backend and

frontend developers, continuous integration knowledge, and security aspects. Often, the complexity of this development is underestimated [90; 91]. ML-pipelines in a narrow sense describe the journey the data takes through pre-processing and feature engineering pipelines, and which model architecture with which hyperparameters is used in detail. Shortly saying, all information that is necessary to reproduce the results. This also includes code and software, as well as dependencies and the programming environment. How well these pipelines are reported was one the research questions of the literature review in section 3.4.

The next subsection addresses the Cross Industry Standard Process for Data Mining (CRISP-DM) and has some correlations with the extended ML pipeline.

2.3.1 | CRISP-DM

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a framework that is also used for ML projects, and we have used this framework for the analysis and communication with stakeholders during this thesis [92]. In contrast to Figure 2.8, it has inner circles that outline the iterative process of ML projects rather than a linear project.

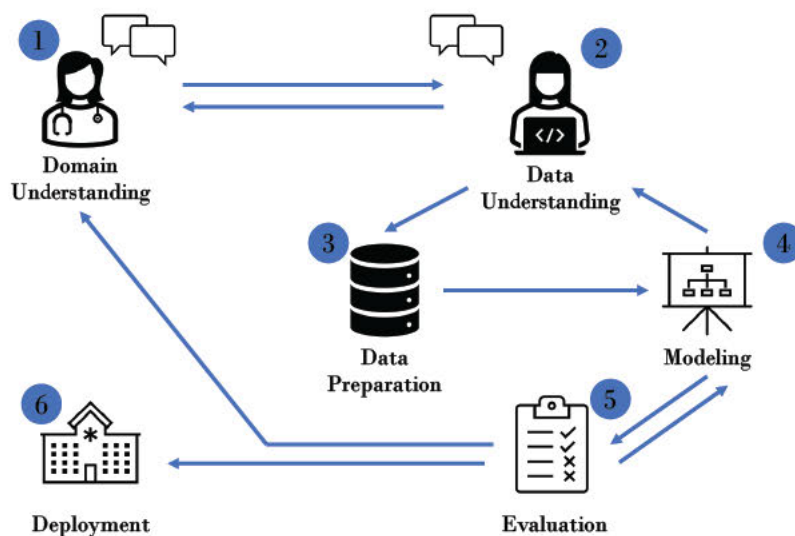


Figure 2.9: Cross-Industry Standard Process for Data Mining (CRISP-DM) cycle. It involves 6 steps from domain and data understanding over data preparation and modeling to evaluation, and deployment.

The CRISP-DM inherits 6 steps: Domain understanding, data understanding, data preparation, modeling, model evaluation, and deployment. Initially, the order of these steps must be kept. However, once one reached the modeling step, it is possible that

the data preparation must be slightly changed to increase model performance, or an increasing data understanding leads to new questions addressed to the domain expert. CRISP-DM therefore has one outer loop and three inner loops. The inner loops are the communication between domain expert and data scientists (1 and 2), the iterative process of developing a model (2, 3, 4), and evaluating a model (4, and 5), the outer loop includes the steps 1 to 5, without deployment. Ending the loop at step 6 is a simplification that is made by CRISP-DM. In a real-world scenario, challenges like concept drift (section 2.4), data drift, and continuous integration must be addressed.

Alternatives to CRISP-DM CRISP-DM is not the only framework that helps to guide through ML projects and pipelines. Verma et. al. propose a 3-phase framework that can be used to describe the implementation and prior development of machine-learned solutions [93]. Phase 1 is the **exploration phase**, where the use case and domain have to be understood, outcomes defined, workflows understood, data feasibility clarified. Phase 2 is the **ML solution design phase**, where a model developed and tested. Phase 3 is the **implementation and evaluation phase** where the solution is implemented iteratively, where a team steps back to phase 1 and 2 before landing to phase 3 again. This three-phases framework has with CRISP-DM in common that it is iterative and includes multiple disciplines like data scientists, software and ML engineers, physicians, project managers, and users. However, there is a larger focus on subject matter experts in the three-phase framework, and CRISP-DM does not include project managers in turn. Also, CRISP-DM ends after deployment whereas this framework contains an iterative deployment.

2.3.2 | Cross-Validation

When at stage 4 (modeling) and 5 (evaluation) of the CRISP-DM cycle, one of the main questions during the project is: How well performs the model after deployment? One of the major challenges of ML projects is the problem of overfitting. A model, that performs very good on train data might have a big performance decrease on test data because it was overfitted. Another one might claim that the test data is not representative for the real-world data and might contain a bias: Too many men, too easy to classify, or only a part of the search space is represented in this dataset. One widely adopted tool to address these challenges is to cross-validate [86]. In cross-validation, one splits up the data into k approximately equal sized folds. In the next step, the model is then train on $k - 1$ folds and validated on the remaining. There exist several ideas which samples (=rows of all available data) should constitute a fold. One can split along a time-axis, along users,

between users, or randomly. Depending on the data and the use case, this can lead to better or worse estimates of the model's performance and generalization ability after deployment. Details about cross-validation are given in section 3.5.

2.3.3 | Scores on Classification and Regression

In an earlier phase of a ML project and when using cross-validation to evaluate your model, there can be several metrics on the evaluation dashboard. For regression tasks, the types of evaluation metrics can be divided into two fields. One field tries to report how far away the model's prediction is from the ground truth. These are metrics like Mean Squared Error, Root Mean Squared Error, Mean Absolute Error, or Median Absolute Error. There is also literature that discusses pros and cons of using one or the other [94]. The other field tries to explain the model's fit on a meta level, like R^2 or the Explained Variance Score. As this thesis mostly train classification models rather than regression, we would like to set the focus more on classification metrics. There are many metrics out there, and sometimes they are confused because of homonyms, i.e., if someone asks "How accurate is your model?" or "How precise is your model?". There are metrics like accuracy and precision that would be the literal answer to these questions, but here can assume that the opponent wants to ask how *good* the model is. And the answer to this question depends on the use case and the aims of the project. One of the best-known metrics is the accuracy score. Given a binary classification task, there are four possible outcomes for a prediction: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). For most clinical datasets, the data is skewed in a sense that there are many healthy patients (target=0) and only very few with a positive disease. (target=1). If the prevalence is 1 % in your dataset, your model has an accuracy of 99 % if you write code that states `print 0`. That's why an accuracy score can be used if the target distribution is balanced in the test set or if the consequences of predicting a FP are the same as predicting a FN . In the medical domain however, this is generally not the case. If your only goal is to detect the positive cases in your dataset, you would optimize the Recall (synonym: sensitivity, true positive rate, hit rate) (TP/P) with P as the number of positive instances. Now again, if you write code that states `print 1` you have successfully optimized your recall, but the model would be obviously useless.

2.3.3.1 | Balanced Accuracy vs F1-Score

Two approaches of handling the precision-recall tradeoff are represented by the balanced accuracy and the F1-Score. The balanced accuracy equally weights true positive and true negative rates: $(TPR + TNR)/2$, with $TPR = TP/P$ and $TNR = TN/N$ with N as the

number of negative instances. The F1-Score, however, does not count for True Negatives at all, meaning it focuses more on the detected positives: $2TP / (2TP + FP + FN)$ which is the harmonic mean of Precision and Recall. Within this thesis, we mostly optimized the F1-Score when training a ML model.

2.3.3.2 | Concordance Index and Brier Score

One of the disadvantages that all previous mentioned classification metrics have in common is that there is a sensitive towards a specific threshold. A classification model predicts not integers but floats, meaning a classification is not 0 but 0.3, for instance. For binary classification tasks, the threshold would be on 0.5, meaning the model predicts 1 if the final value is 0.5 or higher. Now if one wants to optimize the recall of a model, he or she can lower the threshold to 0.4. This shifts the distribution of predictions towards more positive cases. Vice versa, one can increase the threshold to 0.6 which shifts the distribution of predictions towards more negative cases. One number that catches the performance of the model when all thresholds are tried on a test set is the area under the receiver operating characteristic curve (AUC-ROC), or Concordance index (C-index or C-statistic). The ROC curve is a graphical representation of the model's true positive rate (sensitivity) against the false positive rate as the classification threshold varies. It is robust to class imbalance and threshold selection and provides a single number. 0.5 means random guessing and 1 is a perfect model. The Brier-Score (BS) is defined as the squared differences between the ground truth Y and the raw prediction value p . It ranges from 0 to 1, with 0 referring to a perfect and 1 to a most poor model. If the prevalence in the dataset is 50 %, a non-informative model's BS is 0.25, however, if the prevalence is only 10 %, the maximum score is 0.09 ($Y(1 - p)^2 + (1 - Y)p^2 = 0.1(1 - 0.1)^2 + (1 - 0.1)0.1^2$) [95]. The brier score is suitable if one wants to assess the goodness of the calibration of the model. For instance, if a model constantly predicts 0.6 at a threshold of 0.5 for positive cases, this would lead to a good F1-score but a bad BS as the distance to Y is 0.4 for each prediction. In this case, the model has a good discrimination but poor calibration. In summary, there is no one metric for optimizing a model. It always depends on the use case, the prevalence in the data, the goals in the project and the consequences of false positive or false negative classifications.

2.4 | Concept Drift

Thinking within the CRISP-DM framework, concept drift is most important at the deployment stage 6. So, if stakeholders agreed on implementing a model in a live environment,

it is very likely that an abrupt concept or data drift occurs. *Concept*, in this context and with our definition of Machine Learning, is a synonym to the word *mapping*. In other words, if the mapping between input A and output B changes, the concept changes. What will happen then is that the performance of the model decreases because the inherently learned concept is no longer valid. Concept drift can occur due to various reasons such as changes in user behaviour, shifts in the underlying distribution of the data, or external factors affecting the data generation process [96]. There are generally two kinds of drift: Steady and abrupt drifts. The latter, i.e., can happen because of an external abrupt event like a lockdown from the government causing customer behaviour to change abruptly. Steady drifts can happen because of degradation of sensor quality, or a disease progression that causes previously made assumption to hold no longer true. One effective way to address concept drift is a continuous integration system with version control over data, the model, and the code. If the performance of the model decreases over time, a threshold can be set to automatically update the model with the latest x samples from the collected data. The model can then be updated with the latest A to B mapping. The consequence of not having a monitoring system that keeps track of the model's performance can be read in recent work of Lyons et. al., and Wong et. al. [97; 98]. The authors showed that a sepsis model widely used in the United States performed far worse than the manufacturer actually claimed. If we now assume that the manufacturer did not report EPIC to have a higher performance than the model ever had from the beginning, we have to assume that the model had an abrupt drift immediately after implementation due to poor external validity, or that the model became worse over time (steady drift), or a combination of both.

Preliminary Summary In this section, so far, we have introduced the basic concepts of supervised ML, major problems like the bias-variance tradeoff, overfitting, and the mainly used algorithm in the papers (tree-based methods). We have differentiated between ML pipelines in a wider and narrow sense, explained the differences of a wider ML pipeline to a CRISP-DM cycle, we introduced the concept of cross-validation, some major classification and regression scores, and explained the difference of steady and abrupt concept drift. For further reading, all these topics are explained in more detail in the references that we cited, and in the methods sections of the papers in the main part of this thesis. We close this chapter with a more detailed introduction into ML explainability with some in depth introduction of common ML explainability methods, their application area, and a quick discussion of pros and cons of each of these methods.

2.5 | Machine Learning Explainability

Machine learning explainability or explainable artificial intelligence (XAI) refers to the ability for humans to understand the predictions made by ML models. In our literature review from section 3.4, we define a ML explainability method as *a method that enables humans to understand why a model makes certain predictions*.

2.5.1 | Taxonomy

ML explainability methods can generally be classified in two categories [99]. The first category is: Is the method **specific** or **agnostic**? Model agnostic methods, which can be applied to any model regardless its architecture, and model specific methods, which are limited to certain model architectures such as neural networks or tree-based methods. The category is: Does the explanation method provide **local** or **global** explanations? Local explanations explain a single predicted instance of a model whereas global explanations generally give insights to feature importances and the model behaviour as a whole. Some authors also bring in a third, complexity-related category to differentiate if the method is *intrinsic* or *post-hoc* interpretable. Classic examples for intrinsic interpretable methods are linear regressions or decision trees, which is a reason for the popularity of these methods. Post-hoc explainability methods are mostly agnostic methods and are applied to random forests, support vector machines, and neural networks, among others. A graphical overview of this concepts is given in the mind map in Figure 2.10.

2.5.2 | Explainability methods

The literature review in section 3.4 gives a more detailed introduction into the taxonomy of ML explainability. This section here, on the other hand, introduces a subset of the most common explainability methods for tabular, image, and text data. The subset was carefully chosen by the amount of Google citations as of December, 23 in 2021, it's open-source ability in either Python or R, or because they solve an issue of a previously published explainability method. In the next paragraphs, each explainability method is dedicated an own paragraph, and within each paragraph, we try to explain the core idea of this explainability method, where it can be applied, its advantages, and limitations. The methods are ordered by their year of publication. As the ML field is highly dynamic and the number of papers published grows exponentially, this list might be outdated by the time you read this. A good introduction to the general topic of XAI is given by

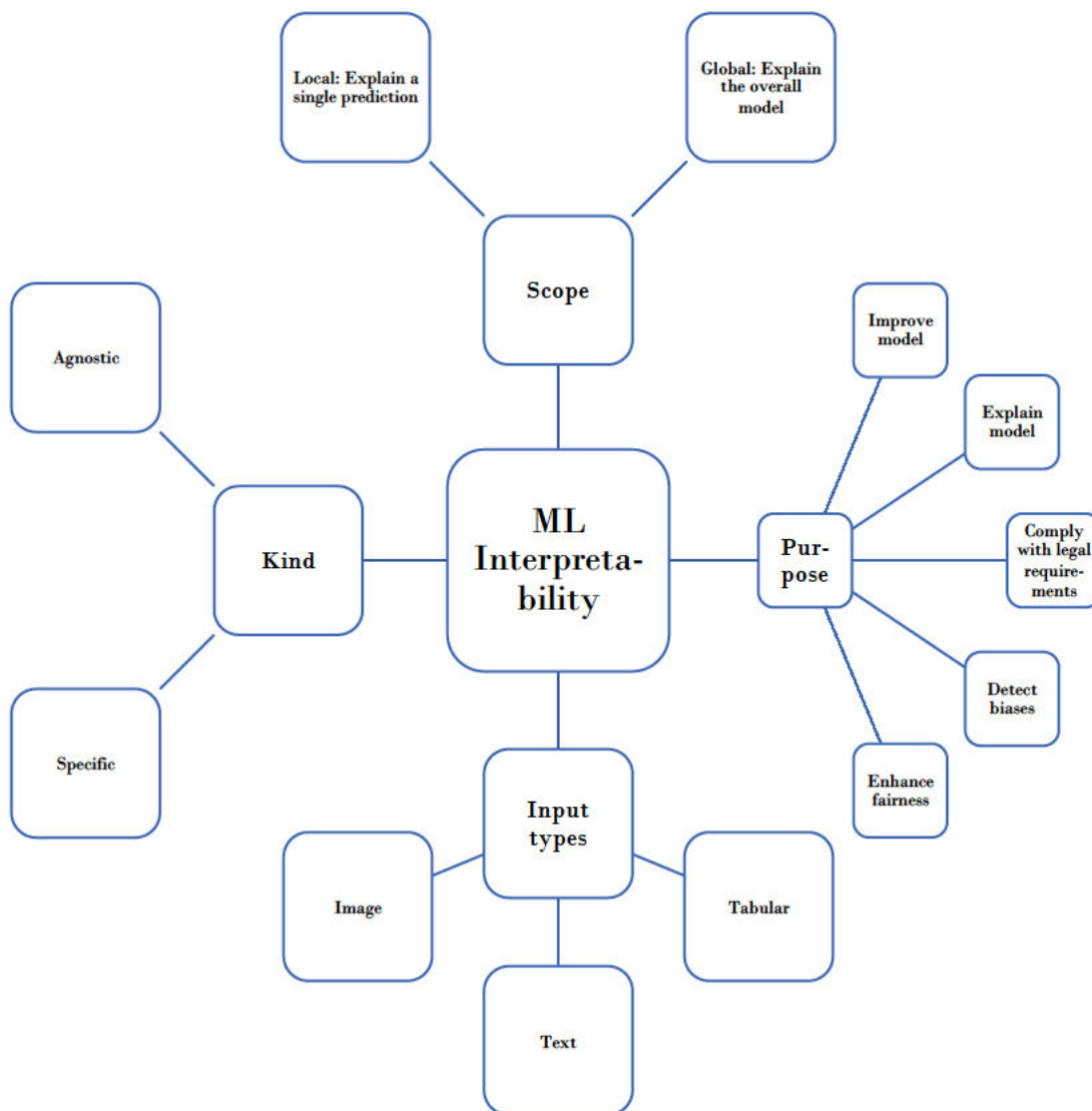


Figure 2.10: Interpretability mind map that provides an overview about the taxonomy of ML interpretability. Note that audio data is out of scope of this mind map and thesis. The reason for that is the literature review did not reveal a single paper that used audio data, ML and XAI in combination.

Molnar et. al [100]. A good overview of the explainability methods given in this chapter here is given in Table 3.17⁴.

⁴This table is a complete copy of my own table from section 3.4 and here to provide a table of content of all the explainability methods that are explained in this section in detail but not in the paper from the results section 3.4 due to length issues.

Method / Taxonomy	Specific (S) or Agnostic (A)	Local (L) or Global (G)	Neural Networks	Computer Vision	Tabular Data	Year	No. Citations	Regr. (R) or Classif. (C)	Source Code Available
Partial Dependence Plots (PDP) [89]	A	G	No	No	Yes	2001	15545	R and C	Yes
Permutation Importance [101]	A	G	No	No	Yes	2010	15545	R and C	Yes
Mean Decrease Impurity [102]	S	G	No	No	Yes	2013	823	R and C	Yes
Individual Conditional Expectation [103]	A	L	Yes	No	Yes	2013	571	R and C	Yes
DeepLIFT (Deep Learning Important Features) [104]	S	L	Yes	Yes	No	2016	1629	C	Yes
Layer-Wise Relevance Propagation [105]	S	L	Yes	Yes	No	2016	2160	C	Yes
Maximum Mean Discrepancy - Critic [106]	A	G	Yes	Yes	No	2016	445	C	Yes
Gradient-weighted Class Activation Mapping [24]	S	L	Yes	Yes	No	2016	6758	C	Yes
Integrated Gradients [107]	S	L	Yes	Yes	No	2017	2017	C	Yes
Local Interpretable Model-agnostic Explanation (LIME) [108]	A	L	Yes	Yes	Yes	2017	5020	R and C	Yes
SHapely Additive exPlanations (SHAP) [109]	A	L and G	Yes	Yes	Yes	2017	5020	R and C	Yes
Leave One Covariate Out [110]	A	L	No	No	Yes	2017	274	R	Yes
Influence Functions [111]	A	L	Yes	Yes	No	2017	1377	C	Yes
Soft Decision Trees [112]	S	G	Yes	No	No	2017	357	C	Yes
SmoothGrad [113]	S	L	Yes	Yes	No	2017	867	C	Yes
Testing Concept Activation Vectors [114]	S	L and G	Yes	Yes	No	2018	583	C	Yes
Anchors [115]	A	L	Yes	Yes	Yes	2018	922	R and C	Yes
Representer Point Selection [116]	S	L	Yes	Yes	No	2018	105	C	Yes
Automatic Concept-based Explanations [117]	S	G	Yes	Yes	No	2019	157	C	Yes

Table 2.1: Overview of interpretability methods relevant to tabular and computer vision tasks, ordered by year of publication. Method relevance to neural networks, computer vision, and tabular data is indicated in the respective columns. The number of citations was derived from Google Scholar as of December 23rd, 2021. Links to the source code are provided via hyperlinks. We included methods that had more than 100 citations on Google Scholar, whose source code was publicly available, and that were optionally used in the review articles. An explanation of each method with advantages and limitations can be found in the supplementary material on [GitHub](#). Regr = Regression, Classif = Classification.

Partial Dependence Plots

Explanation Partial dependence plots (PDP) is a global, model agnostic method that shows the marginal effect of a feature on the target. In principle, PDP answers the question *"What is the relation of these features to the target given that other features are held constant?"*. **Application** The plots are applicable on any type of features (categorical and continuous) as well as to classification and regression problems. **Advantages** Since this method is detached from the ML model, one is independent of the algorithm and primarily looks at the data itself. An implementation can be seen [here](#). **Limitation** However, PDP makes the naive assumption that features have no correlation with each other. Averaging many data points further results in loss of information regarding the heterogeneity of features.

Permutation importance

Explanation Permutation importance is a model agnostic and global method to estimate how important a feature for a given trained model is [101]. It can be used for both regression and classification tasks. Permutation Importance is defined as the absolute difference in the performance score when a feature is replaced by a dummy feature. The more the performance drops, the more important this feature is for the model. A little-noticed variance of permutation importance is the **perturbation rank**, in which the values within a feature are shuffled [118]. The advantage of this method is that statistical properties of feature and dummy feature remain identical. **Application** For Python users, there is an implementation of the method in [scikit-learn](#). The method can be used whenever tabular data is used, regardless of the model. **Advantages** An advantage of the method is its intuitive comprehensibility to stakeholders and the open-source implementation by scikit-learn. **Limitation** A disadvantage is that the Permutation Importance depends on the model and the selected performance score. A modification of the performance scores can mean a change in the feature rankings. In addition, this method cannot take into account co-variances between features.

Mean Decrease Impurity

Explanation Mean Decrease Impurity (MDI) is a global and model specific method for explaining feature importance of ensembles of trees [102]. It aims to identify irrelevant features for the target and attributes a relevance to each feature. For each feature, it calculates the importance (i.e., Gini or Shannon [119]) as the sum over the number of splits for all trees with that feature, proportionally the total number of samples it splits. **Application** Conceivable applications are classification tasks with categorical variables as input, although according to the authors, regression tasks are also conceivable if

one varies the method used to measure impurity. **Advantages** The methodology is mathematically sound within the paper. **Limitation** However within the paper, MDI is shown using categorical input and output variables, which limits potential use cases. At the same time, on the [documentation side](#) of permutation importance, scikit-learn criticizes a bias of MDI towards categorical variables. Perturbation Ranking, introduced by [Jeaff Heaton](#) shuffles the values within one feature, however, the statistical property of the feature (min, max, mean, std) remains the same.

Individual Conditional Expectation

Explanation Individual Conditional Expectations (ICE) are a refinement of Partial Dependence Plots and address the heterogeneity of individual data points [103]. It is a local, model agnostic and post-hoc method which simply disaggregates PDPs to shed light on individual conditional expectations. **Application** It can be applied to supervised applications and is essentially used to illustrate up to 3 variables. There is an implementation in [R](#) and [Python](#). **Advantages** In classical PDPs, averaging results in a loss of information. Disaggregation can reverse this loss of information. **Limitation** ICEs can become confusing if there are too many heterogeneous data points. Although this shows the heterogeneity of the problem, it also makes it difficult to derive concise statements.

DeepLIFT

Explanation Deep Learning Important FeaTures [104] is a local and model specific method using to explain individual predictions of neural networks. It can also be used for computer vision applications. DeepLIFT looks at how much each neuron in a neural network is activated relative to a reference input for an individual input. The reference input is neutral *foil*, whereas the individual input can be described as *fact*. **Application** In the paper, DeepLIFT is applied to MNIST dataset and for classification of DNA sequences. **Advantages** Other methods like [120; 121; 122] also need a forward propagation to for each perturbation and might therefore be computationally inefficient. **Limitation** DeepLift itself has the limitation that it is difficult to generate a suitable reference input (foil) from the data to explain the individual input relative to the reference input.

Layer-Wise Relevance Propagation

Explanation Layer-Wise Relevance Propagation (LRP) [105] produces relevance scores for the input pixels by iteratively distributing the final score across the neural network's layers, starting from the output layer and proceeding backwards to the input layer. Values greater than zero indicate that a particular pixel is relevant for the chosen class. There are several variants of LRP; while LRP was not originally described as a gradient-based

explanation method, it was later shown [123] that ϵ -LRP is a variant of the Gradient * Input method in which the gradient calculation is modified based on the ratio between the output and input at each nonlinearity. **Application** LRP has been applied to image classification models, bag-of-words models [124], and Fisher Vectors [125].

Maximum Mean Discrepancy - Critic

Explanation Maximum Mean Discrepancy *MMD-Critic* is a global model agnostic method that distinguishes representative samples of a class from outliers [106]. Typical representative samples are called prototypes, the outliers are called criticism. Samples in a distribution with high sample density are seen as good prototypes. By detecting the criticisms, a higher interpretability of black-box models should be achieved. **Application** Typical applications are examples of computer vision models, but tabular applications are also conceivable. The source code is freely available on [GitHub](#). **Advantages** MMD-Critic works for any data type and any model. It is therefore maximally flexible. It can help people when labeling images to recognize untypical images of a class more reliably. **Limitation** A criticism is not necessarily harder to classify from the model. As an alternative, a classical error analysis of misclassified samples can help to detect difficult samples systematically by examining these images for concepts.

Gradient-weighted Class Activation Mapping (Grad-CAM) and Guided Grad-CAM

Explanation Gradient-weighted Class Activation Mapping (Grad-CAM) is a model specific, local and post-hoc explainability method for computer vision tasks and reinforcement learning [24]. It calculates a linear combination of neuron importance weights and feature map activations for the last convolutional layer as this layer has the best compromise between spatial information and high-level semantics. Grad-CAM basically answers the question: Which part of an image is important for a specific classification? **Application** Grad-CAM can be used to explain computer vision models solving object detection, image classification and visual question answering tasks. The model has been evaluated on datasets like [ImageNet](#), [COCO](#), [Visual Question Answering](#) and [Places](#). The source code is freely available on [GitHub](#). **Advantages** Although we categorize the method as model specific, it is applicable to a variety of CNN model families such as fully connected ones, multi-modal inputs for visual question answering or structured outputs such as captioning. Unlike the older CAM algorithm [126], the Grad-CAM method is a generalization in that it does not require a specific model architecture. **Limitation** A potential limitation could be the "Guided Grad-Cam" variation presented in the paper. Guided back-propagation acts more like an edge detector than providing insights into

the model behavior [127; 128]. Solutions for this could lie in further developments of CAM, such as Grad-CAM++[129].

Integrated Gradients

Explanation Integrated gradients calculates step by step the difference of a neutral input (a baseline, i.e., a black image) to a given input [107]. The gradient provides an estimator of which value weights most strongly for prediction. **Application** Integrated gradients was demonstrated by the authors on image models, text models, and a chemistry model. The method has even the ability to debug a model. **Advantages** The method does not require any modification to the model of interest and can directly be applied to the standard gradient operator. The paper presents two axioms that, according to the authors, should be fulfilled for an attribution method, namely *sensitivity* and *implementation invariance*. Sensitivity is given when a different feature between input and baseline is non-zero. Implementation invariance is given if a neural network always outputs the same prediction for a given input, regardless of its architecture. According to the authors, DeepLIFT, i.e., breaks both of these axioms. Integrated gradients can be applied to any differentiable model. **Limitation** The limitation of this (and other image attribution methods) is that interactions between features as well as the logic of the network are not addressed.

Local Interpretable Model agnostic Explanation

Explanation Local Interpretable Model agnostic Explanation (LIME) is a popular, open-source, and post-hoc method that learns an interpretable model around a single prediction. Using data points close to the individual predictions, LIME trains an interpretable model to approximate the predictions of the real model. The new interpretable model is then used to interpret the result, which is also called local fidelity. **Application** LIME can locally explain text-models from tree-based algorithms as well as computer-vision models, such as deep neural networks. **Advantages** LIME breaks the complexity of a global model by taking samples that are locally close to a prediction. **Limitation** It has been shown that a random generation of noise results in an instability of the generated explanations by LIME [130; 131]. This results in modifications of the originally posted LIME approach, i.e., S-LIME [132] or DLIME [131].

Shapley Additive exPlanations

Explanation SHapley Additive exPlanations (SHAP) is a model agnostic method that allows both global and local explanations and also addresses structured as well as unstructured data [109]. SHAP is the contribution of a feature value to the difference

between the actual prediction and the mean prediction. The popularity of SHAP is not least explained by the freely available source code on [GitHub](#). SHAP builds on Shapley values from game theory [133], propagation activation features [134], and model-intrinsic approaches from [tree-based methods](#), among others. **Application** SHAP is written in Python and can be applied to models from Tensorflow, Keras, Pytorch and scikit-learn. The built-in visualization functions facilitate the interpretation of the methods. **Advantages** By combining different methods on a high-level API, SHAP is also available to a wider audience. This is a decisive advantage over other methods. **Limitation** However, there is also the danger that SHAP is applied without questioning the limitations of the underlying methods.

Leave One Covariate Out

Explanation Leave One Covariate Out (LOCO) is a model agnostic, global and local feature importance method similar to feature importance in random forests [110]. In contrast to feature importance in random forests, however, the feature under consideration is not replaced by a dummy variable, but simply dropped. **Application** The authors themselves describe regressions as a use case, although classification is also conceivable. There is also a [GitHub](#) repository freely available in R. **Advantages** One advantage of this method is its simple implementation. Although there is a GitHub repository for R, you can also implement LOCO yourself with a for loop. **Limitation** It remains unclear whether LOCO offers a real advantage over Breimann's older feature importance.

Influence Functions

Explanation Influence Functions is a local, model agnostic explainability method for providing training points most responsible for a given test sample [111]. An Influence Functions treats the model as a function of the training data. It gives more weight to a single sample and examines the change in output when that sample is changed. **Application** In the paper, the Influence Functions are applied to animal images. The authors also show the outlier sensitivity of a model when noise is applied to important images and added to the training set. The source code is available on [GitHub](#). **Advantages** The method can be applied to all machine learning models whose 2nd degree derivative exists. **Limitation** However, the method is very computationally expensive because the model must be re-trained when the training data changes. In addition, the boundary of an influencing or non-influencing training example is unclear.

Soft Decision Trees

Explanation Soft Decision Trees is a model specific and global interpretability method

which uses a decision tree to mimic the input-output function of a neural network [112]. In a soft decision tree, all the leaf nodes contribute to the final decision with different probabilities [135]. **Application** The authors demonstrate the soft decision tree using the MNIST dataset. Inner nodes of a soft decision tree represent learned filters of the neural network. The code was re-implemented by third parties on [GitHub](#). **Advantages** For some leaf nodes, the soft decision tree allows the visual interpretation of the neural network. The simplification of the complex network architecture results in a leaner model with relatively low performance loss. **Limitation** Not all learned filters are interpretable to the human eye. The explainability of this method is therefore limited.

SmoothGrad

Explanation SmoothGrad is a model specific, local and post-hoc explainability method that tries to reduce noise in saliency maps (also called sensitivity maps or pixel attribution maps) for model explanation [113]. In the neighborhood of an input image x , random examples are generated and blended with the sensitivity map by averaging. **Application** The authors apply the method to their own input images and parts of the MNIST dataset. The source code is freely available on [GitHub](#). **Advantages** For some input images this method works better than comparable ones like Integrated Gradients [136] or Guided Backpropagation [137]. The method can also be combined with other methods. **Limitation** It remains unclear for which type of images the method works better than others. There is no discernible pattern for the examples shown in the paper.

Testing Concept Activation Vectors

Explanation Testing Concept Activation Vectors (TCAV) is a model specific, global and local explainability method for computer vision models and tabular, discrete data [114]. TCAV gives an explanation (i.e., a concept) that generally applies to a class which is beyond one image. It learns the concept from examples. The concepts are learned through delineation examples. For example, to learn the concept feminine, some images of feminine must be shown in differentiation from non-feminine. **Application** The source code is freely available on [GitHub](#). In order to apply TCAV, two data sets must be provided. One representing the concept and a random dataset for delineation. We then train a binary classifier to distinguish between the concept and the random data. The coefficient vector of the classifier is then called a concept vector. **Advantages** Since people think in concepts and not in numbers, this method is also applicable for non machine learning experts. **Limitation** The concept datasets need additional labels and therefore could be expensive to create. Also, abstract (i.e., sadness) or too general concepts are difficult to learn.

Anchors

Explanation Anchors is a model agnostic local explanation method developed by the LIME authors [115]. Based on a prediction, relevant features are determined. If a marginal change in other features does not change the prediction, then the rule is *anchored*. The outputs of the anchors approach are IF-THEN rules. **Application** Anchors can be applied to structured predictions, tabular classification, image classification, and visual question answering. The source code is freely available in [Python](#) and [Java](#). **Advantages** By generating if-then rules, the output of this explanation method is easy to understand even for non machine learning experts. In addition, Anchor offers a very wide range of applications. **Limitation** Rules for rare classes or near the boundary of decision functions can become complex and sometimes ambiguous. With complex output, different rules can also become the same prediction. In high dimensional spaces also every small change can lead to a change of the prediction, which makes the coverage of the rule very low.

Representer Point Selection

Explanation Representer Point Selection is a model specific, local explainability method for computer vision applications [116]. For a given test image, *representer points* are similar images from the training set and are close to the decision boundary. Positive representer points belong to the same class as the test image, negative ones to a different class. **Application** The method can be applied to any image classification task. The source code is available on [GitHub](#). **Advantages** By showing these images, the method helps in error analysis and model understanding. Within the paper, examples are also shown which demonstrate an improvement to the Influence Functions method. **Limitation** It is questionable whether the method has advantages over a classical error analysis. By displaying the misclassified images, one can look for systematic errors by clustering them. This is in essence also what Influence Functions and Representer Point Selections do.

Automatic Concept-Based Explanations

Explanation Automatic Concept-Based Explanations (ACE) is a global, model specific interpretability method to cluster and visualize segments of an image that are important for a particular class [117]. Multiple images for one class are segmented using the activation space of a layer of a pre-trained neural network as a similarity score. Similar segments of the images are then pooled together. Finally, each pool is assigned a TCAV importance score. **Application** The method can be applied for any computer vision classification model. The source code is written in Python and available on [GitHub](#). **Advantages** By pooling multiple images, this method can also be considered as a global

explanation method for computer vision use cases. **Limitations** The explanation method only works if the concepts are present in the form of groups of pixels. Abstract concepts do not work.

In this last section 2.5 of the materials and methods chapter, we tried to give an overview of commonly known explainability methods. You will see in the main part (section 3.4) that in the medical domain mainly SHAP and Grad-Cam are used, because they are well known and generically applicable. In the discussion of the literature review and in the discussion section of this thesis, we also address the limitations of these explainability methods.

Results

This cumulative dissertation follows a structured arrangement where the thesis contributing papers are presented in a sequential and thematic order. The arrangement of these papers is determined both by their chronological progression and the nature of their content. The focal point of the research revolves around the medical domain of tinnitus, which is primarily investigated in sections 3.1 and 3.2. Another domain explored is the coronavirus, discussed comprehensively in section 3.3. The exposition begins with a comprehensive literature review concerning machine learning's interpretability within the medical domain. This review serves as the foundation for the subsequent contribution, which provides a taxonomy of explainable artificial intelligence (XAI) methods, expounded upon in section 3.4. The chapter closes in the amalgamation of all datasets involved in this thesis, explored comprehensively in section 3.5. This final paper scrutinizes the intricate interplay between the challenges posed by EMA and MCS data and their subsequent impact on the accuracy of machine learning performance estimation. It is pertinent to note that not every research question is exhaustively addressed in every section, as each section is dedicated to a specific aspect of the overall research endeavor. To facilitate seamless navigation and comprehension, a mapping between each research question and its corresponding paper is provided.

- Main RQ1 (*How can machine learning help confirming or broaden domain knowledge within mHealth data?*) is addressed in section 3.1, section 3.2, and section 3.5.
- Main RQ2 (*How can one reach explainability in the presence of mHealth data when using Machine Learning?*) is addressed in section 3.1, section 3.2, and section 3.4, and
- Main RQ3 (*Which guidelines can be beneficial for the use of ML within the mHealth domain?*) is addressed in section 3.4, and 3.5.

3.1 | Predicting the Gender of Individuals with Tinnitus based on Daily Life Data of the TrackYourTinnitus mHealth Platform

- **Authors** | Allgaier, Johannes; Schlee, Winfried; Langguth, Berthold; Probst, Thomas; and Pryss, Rüdiger
- **Published in** | Nature Scientific Reports, 11(1), 18375, 2021.
- **Available at** | <https://www.nature.com/articles/s41598-021-96731-8>

Abstract

Tinnitus is an auditory phantom perception in the absence of an external sound stimulation. People with tinnitus often report severe constraints in their daily life. Interestingly, indications exist on gender differences between women and men both in the symptom profile as well as in the response to specific tinnitus treatments. In this paper, data of the TrackYourTinnitus platform (TYT) were analyzed to investigate whether the gender of users can be predicted. In general, the TYT mobile Health crowdsensing platform was developed to demystify the daily and momentary variations of tinnitus symptoms over time. The goal of the presented investigation is a better understanding of gender-related differences in the symptom profiles of users from TYT. Based on two questionnaires of TYT, four machine learning based classifiers were trained and analyzed. With respect to the provided daily answers, the gender of TYT users can be predicted with an accuracy of 81.7%. In this context, worries, difficulties in concentration, and irritability towards the family are the three most important characteristics for predicting the gender. Note that in contrast to existing studies on TYT, daily answers to the worst symptom question were firstly investigated in more detail. It was found that results of this question significantly contribute to the prediction of the gender of TYT users. Overall, our findings indicate gender-related differences in tinnitus and tinnitus-related symptoms. Based on evidence that gender impacts the development of tinnitus, the gathered insights can be considered relevant and justify further investigations in this direction.

3.1.1 | Introduction

Many people experience a long-term noise in their ears, which is widely known as tinnitus, also described as a whistling or ringing sound [138] in the ears. About 10 - 15% of the worldwide population report this kind of symptoms [139; 140]. Although

many people perceiving tinnitus do not experience a considerable burden, about 2.4% of the worldwide population severely suffers from tinnitus on a daily basis [141]. In most of these cases, tinnitus is a subjective perception that can only be perceived by the affected person. Inversely, rare forms of tinnitus exist, for which the perceived sound is caused by a source in the body that can be objectively measured (e.g., blood flow or muscle contractions). As an important consequence of the discussed aspects, no general treatment, which is able to effectively reduce tinnitus symptoms like loudness and its related fluctuation, exists yet. On the individual basis, tinnitus can be reduced, for example, by the use of cognitive behavioral therapies [142]. To characterize the general status of available treatments with respect to the well-known heterogeneity of tinnitus patients [143; 144], they are rare and their development is difficult.

To better and more effectively deal with this heterogeneity, researchers often focus on the identification of subgroups of tinnitus patients. Identified subgroups might be used for investigations on treatments for an identified subgroup instead of a general treatment for all tinnitus patients. However, the clustering of tinnitus patients through the identification of subgroups is not an entirely new research question. Hitherto, several approaches aimed at the clustering of tinnitus patients depending on their symptom profiles [145; 146], or depending on neuroimaging data [147]. Furthermore, the authors of [148] developed the Tinnitus Primary Function Questionnaire to examine the effect of tinnitus on thoughts and emotions, hearing, sleep, and concentration. The authors established correlations between these four effects and derived secondary limitations for the individuals in their daily life. The consideration of potential differences in gender are another approach on subgroup research. A recent special issue shows the latter kind of interest in research [149]. In the already published articles of this special issue, for example, one work deals with gender differences of chronic tinnitus patients [150]. All of the presented works show that gender differences are a valuable research direction in particular and with respect to research on subgroups of tinnitus patients in general. In addition, research evidence exists that the gender impacts the development of tinnitus and the response to treatments. For example, in this recent work [151], the authors investigated treatments of 316 patients and found significant treatment differences between males and females. For instance, females improved better in orofacial therapies. Or, in the work of [152], it was found, among other findings, that stress was positively correlated with tinnitus severity only in males. These and other findings clearly show that gender-related differences are relevant for investigations of tinnitus patients and their symptom profiles.

In the discussed context, the use of mobile applications to monitor health symptoms is becoming more and more popular, also denoted by mobile and digital health (mHealth).

With respective mHealth solutions, the collection of data becomes easily possible, especially on a daily basis. Furthermore, data can be collected close to the user's daily life with the goal to foster self-monitoring and eventually may support health care in clinical practice [153]. For example, the authors of [154] monitored and investigated mental health conditions by using a mHealth solution, while the authors of [10] showed the general potential and impact of mHealth applications. For TrackYourTinnitus (TYT), the daily use, among other reasons, enables individuals to better deal with the variations of the tinnitus over time. On the flip side, mHealth solutions also revealed drawbacks, which are discussed by many recent works. For example, potential discrepancies of app developers and patients of mHealth apps are investigated more in-depth by [155], while general challenges are discussed by [156]. In the discussed setting, it should always be kept in mind that a daily smartphone usage might also worsen the individual tinnitus situation as users are reminded about their problems on a frequent basis. However, research works exist that have shown that the daily use of mobile technology does not aggravate the overall health condition, see for example [157]. Despite such findings, the daily focus on a disease when using mHealth solutions should always be considered carefully.

For the identification of tinnitus subgroups, the collection of longitudinal ecologically valid data sets based on mHealth solutions has been recognized by several researchers. Technically, mobile crowdsensing techniques [158] or Ecological Momentary Assessments [30] are mainly utilized to gather the required data sets. For tinnitus research, these technologies have already shown that they can collect valuable data [159; 160]. To identify subgroups of tinnitus patients, data sources established by the use of mHealth solutions have also revealed to be appropriate [161]. Several of these works have presented their findings on data of the TrackYourTinnitus platform (TYT), which was developed to evaluate daily symptom fluctuations of tinnitus patients. TYT comprises two mobile native (developed without using frameworks) applications (an Android and an iOS app), a website (www.TrackYourTinnitus.org), and a server application that stores the data generated by the apps. The platform was developed by an interdisciplinary team of computer scientists, medical doctors, and psychologists. It can be freely used by interested users, the apps can be downloaded through the official app stores from Apple and Google. In essence, the following complete the following procedure: First, they have to fill out three registration questionnaires after downloading the app. After that, they decide on the number of daily notifications. Each notification reminds the user to fill out a daily questionnaire, comprising so-called EMA questions, which aim at the momentary tinnitus situation of a user. In addition, the environmental sound level is collected through the microphone of the used smartphone when filling out the daily questionnaire.

In terms of feedback, the app visualizes the gathered data and through the website, interested users can download their collected data. TYT does not offer further features. Although the platform aims at data for research and it could be assumed that this is of less interest, so far, the platform has gathered more than 100,000 daily questionnaires by more than 3,000 users from all over the world. We learned that despite the fact that TYT is an open research project in the sense of a long-running observational study, two aspects are of importance for users to participate. First, the project is without any commercial interest. Second, data is collected anonymously except one reason. If users want to reset their password, they have to provide their mail address. In general, the secure handling of data collected by the use of a smartphone is an important aspect since smartphones provide a lot of opportunities to gather data that indirectly might reveal the user. For example, when GPS data is collected and the location of a user is sent to a central server. In general, works exist that have developed complex configurations with which users can control the provision of mHealth-related, see for example [162]. Interestingly, such works show that users are less interested to control much themselves, therefore it is important that a mHealth solutions tries to secure data and privacy in the best possible way by design. In the case of TYT, only questionnaire data and the environmental sound level are gathered, which might be also one reason to use it frequently by many users. To conclude, the TYT project is running since 2014 and revealed various investigation opportunities, including those, which were initially not planned [161; 163]. Beyond TYT, other mHealth solutions have been developed and presented to support diagnosis and therapy of tinnitus patients [142; 164; 165], which emphasizes the potential of mHealth in this context.

Moreover, the combination of mHealth and machine learning has become very popular recently. The directions followed in this context are manifold. On the one hand, considerations on sparse mHealth data are subject to research when using machine learning methods in the given context [166; 167]. On the other hand, large mHealth data sets exist that are investigated by the use of machine learning methods [168]. Moreover, the development of new machine learning methods and the evaluation of existing ones is also considered presently [169; 170].

In this work, gender-related differences of TYT users are investigated, hereby based on the following thoughts: Existing insights on TYT, existing works on machine learning methods to identify subgroups of TYT users, and the amount of existing data of TYT users distributed between females and males. Further note that TYT is technically based on mobile crowdsensing techniques [4] and utilizes Ecological Momentary Assessments (EMA) to capture ecologically valid data sets of tinnitus patients. Since 2014, the TYT mHealth platform has gathered more than 100,000 completed questionnaires from its

users. With respect to the identification of subgroups, machine learning based investigations on the TYT source already exist. For example, in [15], the differences of TYT Android and iOS users were investigated, while in [171], entity (i.e., individual TYT users) similarity was investigated to label the future observations referring to an entity. For the investigation at hand, two prerequisites are important: First, it must be defined which type of gender differences are addressed in this work. The authors of [149] define the following important differences: the (1) biological classification encoded in the DNA and the (2) understanding of the respective social roles, behavior, and expressions. In this work, we refer our considerations to the latter type of difference. Second, it must be defined which gender-related aspects of TYT users shall be investigated. The answer to this question is that our goal is to predict the gender of the user of a provided daily assessment. A daily TYT assessment, in turn, is based on the filled-out daily questionnaire, which comprises 8 EMA questions (users can opt which questions they actually want to fill out; in addition, 1 question varies among users based on an answer given to the perceived worst symptom provided through one baseline questionnaire) that capture the current situation of a TYT user (see this work for a detailed explanation [172]). Note that TYT users have two options to fill out this questionnaire. The first option entails receiving up to 12 random notifications per day, which then remind users to fill out the questionnaire, while the second option allows users to determine fixed points in time to receive the notifications. Furthermore, baseline questionnaires, which must be answered when using the smartphone app for the first time, provide the information on the gender of a TYT user. Based on this information, 15 features were identified - out of the 8 daily questions - for the gender prediction task, covering aspects like stress, worries, arousal, depression, mood, or the loudness of the momentarily perceived tinnitus. A detailed explanation of the features is provided in Table 3.3.

Given these two prerequisites, the overall goal of the work at hand is the prediction of the gender of the user of a given daily TYT assessment based on machine learning methods. A binary classification is therefore accomplished that deals with the following detailed questions (note that for the classification task, technically, scikit-learn[173] has been used):

- i Is it possible to learn a mapping function from X to y of TYT individuals, for which X are questions that the user answered daily, and y is a binary target representing the gender of the respective TYT user?
- ii Which machine learning model is mostly suitable for this task and has a high prediction power?

iii Which are the features with the highest importance to predict the gender?

It is briefly discussed whether other approaches have trained binary classifiers on mHealth related data with respect to research questions on gender-related differences. In general, works exist that have trained a binary classifier on mHealth data. For example, the authors of [174] used such a classifier for respiration disorders of mHealth applications. Furthermore, approaches exist that investigated gender differences in the general context of mHealth solutions. However, their focus is different to the one that is investigated in this work. More specifically, other works [175; 176] investigate differences when using mHealth technologies from a general point of view. That means that they investigate whether there is a difference between men and women when addressing medical issues while using mHealth solutions. Yet, the focus of these works is different to the presented work: they start with the gender and try to establish which bias this might generate on the use of a solution. In contrast, this work starts from the data source and tries to predict the gender. Although these two perspectives address the same overall research context and are therefore intertwined, the research questions they are addressing are different. Still, to the best of the authors' knowledge, similar works that present a binary classifier on mHealth data with respect to results on gender-related differences do not exist yet.

3.1.2 | Results

In this section, the three research questions are discussed subsequently. First, it is discussed whether it is generally possible to solve the gender prediction task by using machine learning with relevant results. Next, the hyper-parameters of the chosen classifiers must be fine-tuned. Finally, by using the knowledge from Research Questions *i* and *ii*, the question must be answered, which of the features are mostly suitable to classify the gender. A summary of this section is provided in Table 3.1.

3.1.2.1 | Research Question i

In this study, gender is considered to be binary as there is no data for diverse tinnitus patients. Given that the target classes are uniformly distributed, random guessing for a binary classification task leads to an accuracy of 50% on average. Consequently, a mapping from X to y is adding information if the accuracy of a classifier is higher than 50%. If it is significantly higher than 50%, it must be decided based on the achieved accuracy whether it is actually relevant or useful. X was used as the (sub)set of features and y as the target for gender, with {male, female} as possible classes.

No.	Research question	Machine learning algorithm				Results
		SVM	Tree	RF	NN	
i	Is it generally possible to learn a mapping function from X to y where X are questions that the user answered daily and y is a binary target representing the gender of a user?	✓	✓	✓	✓	Precision on average: Male: 81.5 % Female: 84.3 %
ii	Which machine learning model is most suitable for this task and a high prediction power?	✓	✓	✓	✓	Mean accuracy on a 5-fold cross validation set: Random Forest classifier (81.7%)
iii	Which are the features with the highest importance to predict the gender?			✓		Most important features are: q8_4: Worries about the tinnitus q8_5: Difficulties in following a conversation

Table 3.1: Overview of the three Research Questions *i-iii*, the used classifiers and the results. SVM = Support Vector Machine, Tree = Decision Tree, RF = Random Forest, NN = Multilayer Perceptron Neural Network. A checkmark means that this classifier has been used to answer the research question.

The classification task was accomplished using Python, as this is one of the most used languages for Machine Learning [177], which enables comparisons to many other research results. Four classifiers from the [scikit-learn](#) library were used for the investigations: A Support Vector Machine, a Multilayer Perceptron Neural Network, a Decision Tree, and a Random Forest. All of them were able to guess the gender with a significantly higher accuracy than 50%. These classifiers were selected as they are well known to get high accuracy scores for high dimensional classification tasks on small to middle-sized datasets [178; 179; 180; 181].

Note that the more features were added to the classifiers, the higher was the accuracy. For the testing set, a 5-fold cross-validation was used to avoid overfitting. As can be seen from Table 3.2, the random forest classifier had the highest prediction power in this distribution.

Classifier	Precision Male	Precision Female	F1-score
Support Vector Machine	0.80	0.86	0.83
Decision Tree	0.81	0.80	0.81
Neural Network	0.82	0.83	0.83
Random Forest	0.83	0.88	0.85

Table 3.2: Comparison of the four used classifiers in terms of precision per gender and F1-score. Number of examples is denoted by $m = 1702$. Used features: $\{q1, q2, \dots, q7, q8_5\}$, test size 20%. Note that the feature labels qx are further explained in Table 3.3.

3.1.2.2 | Research Question ii

As there is no other satisfying metric such as training time or minimal false positives rates, it was decided to further investigate the classifiers accuracy.

To do so, a fine-tuning of the hyper-parameters of the Random Forest classifier was performed. This tuning is also known as a grid search [182; 183]. Therefore, the hyper-parameters of interest were selected, which can be seen in Fig. 3.1. Then, one of the hyper-parameters was varied while keeping all others constant. The resulting parameters-dictionary was passed to the Random Forest classifier into the same training and testing set of the approaches of Research Question *iii*, again with a 5-fold cross-validation [184; 185; 186]. Here, a 5-fold split was used instead of a 10-fold split for the purpose of having a sufficient testing size. Additionally, this allows to speed up training and testing time as well as to vary more hyper-parameters within the grid search. The cross-validation further prevents the Random Forest from overfitting of the training set [60]. For each possible combination of the parameters dictionary, the accuracy was saved. After trying all variations, the variation with the highest accuracy determined the final parameters set up of the Random Forest classifier in the testing set.

```
parameters = {'bootstrap': [True, False],
              'ccp_alpha': [0.0],
              'class_weight': [None],
              'criterion': ['gini', 'entropy'],
              'max_depth': [None, 2, 5, 10, 20, 100],
              'max_features': ['auto', 'sqrt', 'log2'],
              'max_leaf_nodes': [None],
              'max_samples': [None],
              'min_impurity_decrease': [0.0],
              'min_impurity_split': [None],
              'min_samples_leaf': [1, 2, 10],
              'min_samples_split': [2],
              'min_weight_fraction_leaf': [0.0],
              'n_estimators': [1, 3, 5, 10, 100, 200,
                               300, 500, 1000],
              'n_jobs': [None],
              'oob_score': [False],
              'random_state': [1994],
              'verbose': [0],
              'warm_start': [True, False]
            }
```

Figure 3.1: Set of hyper-parameters for a grid search in order to improve the forest's accuracy. Note that not all hyper-parameters have been varied, such as `n_jobs`, `oob_score` or `verbose`. Only hyper-parameters were varied that have a higher impact on the accuracy score. However, static parameters are listed for the purpose of integrity.

The number of decision trees in the random forest was increased up to 1,000 for a slight improvement of the overall accuracy. However, a further increase of `n_estimators` did not improve the score in the testing set. If the `max_depth` parameter was lowered to 10, the lowest standard deviation of 2% within the 5-fold cross-validation was attained.

The best ranked Random Forest classifier received an accuracy of 87% in the first cross-validation set. The average cross-validated test score is **81.65%**, with a standard deviation of 4%.

3.1.2.3 | Research Question iii

There exist several techniques to determine feature importance, such as random, heuristic, or complete approaches [187]. In order to answer the third Research Question iii, three strategies were pursued. Before the strategies were accomplished, a sub-dataframe was created that contains the feature of interest and the target gender. This sub-dataframe was then filtered, so that it equally contains 50% men and 50% women.

As the first strategy, a closer look was put on the random forest approach. Importantly, it has no bias in terms of the underlying distribution of the mapping function. The forest simply measures the impact in accuracy. The higher the accuracy score for a mapping from a feature to the target is, the higher its impact on the target is. The second approach tried to measure the impact of single features using correlations with the target gender. The correlation matrix also helps the authors to get a more detailed insight into the cross-correlation between the features and a single-viewed impact of a feature on the target. The higher the correlation is, the higher the impact to the target is. Note that the correlation method varied with the scaling (binary, discrete, continuous) of a feature. For a univariate classification on gender, a rise in accuracy was expected if the correlation rises. Third, the permutation importance for a univariate Random Forest classification per feature was calculated [188] as follows: First, the classifier was trained on a training set. Then, using cross-validation, a baseline metric was evaluated on a testing set. The permutation importance was then defined as the difference of the baseline metric with the trained feature and the baseline metric with a completely random, artificial feature. All approaches have different units to measure the impact (Accuracy, r-value, and percentage improvement). In order to make these three approaches comparable, a ranking of the results of the three approaches was created (see Fig. 3.2), and statistics for the two gender groups added, respectively. The dynamic questions q_i , with $i = 0, 1, \dots, 8$ have on average a better ranking than the questions q_1, q_2, \dots, q_7 . Throughout all three approaches, *strong worries* (ranked first) and *difficulties in following a conversation* (ranked second) are the two most important features in order to predict the gender. The p-value column shows that these gender differences are all significant. From a statistical point of view, the mean difference between the two groups *male* and *female* generally supports the hypothesis that male individuals experience tinnitus differently than female individuals.

Features		Univariate feature ranking			Statistics								
Question	Label	Correlation	RF Importance	Permutation Importance	Mean male	Mean female	Mean diff.	Std. male	Std. female	T-test	P value	Effect size	
Did you perceive the tinnitus right now?	question1	9	13	11	0.79	0.69	0.10	0.41	0.46	t(20692) = -23.46	<.001	0.23	
How loud is the tinnitus right now?	question2	14	6	5	0.46	0.46	0.01	0.30	0.28	t(20692) = -1.82	0.068	0.02	
How stressful is the tinnitus right now?	question3	15	12	10	0.35	0.36	-0.01	0.28	0.27	t(20692) = 2.66	0.008	-0.03	
How is your mood right now?	question4	6	8	8	0.56	0.57	-0.01	0.21	0.22	t(20692) = 5.57	<.001	-0.05	
How is your arousal right now?	question5	8	14	9	0.25	0.29	-0.03	0.22	0.23	t(20692) = 15.46	<.001	-0.15	
Do you feel stressed right now?	question6	11	9	7	0.27	0.30	-0.03	0.24	0.23	t(20692) = 12.29	<.001	-0.12	
... concentrate on the things you are doing right now?	question7	12	5	3	0.58	0.60	-0.03	0.31	0.30	t(20692) = 8.34	<.001	-0.08	
... it is hard for me to get to sleep.	question8_0	7	10	12	0.38	0.29	0.09	0.49	0.45	t(2461) = -7.43	<.001	0.21	
I am feeling depressed...	question8_1	10	11	15	0.21	0.22	-0.01	0.41	0.47	t(1399) = 7.62	<.001	-0.25	
I find it harder to relax...	question8_2	13	15	14	0.44	0.46	-0.02	0.50	0.50	t(4578) = 2.52	0.012	-0.05	
I have strong worries...	question8_4	1	1	1	0.26	0.43	-0.17	0.44	0.50	t(1023) = 9.45	<.001	-0.35	
...difficult to follow a conversation...	question8_5	2	2	2	0.39	0.22	0.17	0.49	0.42	t(4253) = -16.65	<.001	0.36	
...difficult to concentrate.	question8_6	5	3	13	0.37	0.54	-0.17	0.48	0.50	t(2211) = 12.15	<.001	-0.37	
...I am more irritable with my family...	question8_7	4	4	4	0.35	0.20	0.15	0.48	0.40	t(694) = -6.30	<.001	0.34	
...I am more sensitive to environmental noises.	question8_8	3	7	6	0.07	0.17	-0.10	0.25	0.37	t(2533) = 10.95	<.001	-0.39	

Figure 3.2: Comparison of three approaches to determine the most important feature for gender prediction. A ranking value of 1 means that this feature is most important to predict the gender.

3.1.3 | Discussion

The authors are aware of the fact that by including the dynamic question q8 (The follow-up questions about the worst tinnitus symptom), only a smaller subset of TYT users could be investigated (out of all individuals), which is predestined to have a higher bias. Instead of 80,966 examples, the subsets had sizes between 3,400 (4%) and 14,000 (17%) user examples. The different sizes of male and female individuals by gender can also be seen in Fig. 3.4. That means., if q8_5 (Difficulties in following a conversation) is chosen, it means that 10.9% of the women are included in the dataset. These subsets decrease in size again once an equal split for the target (50% men and 50% women) is performed. As a conceivable result, these subsets could not be representative anymore for the underlying distribution that has a size of $m = 80,966$. Consequently, the distribution of the chosen subset was compared with and without feature q8_5 (Difficulties in following a conversation). Note that the features q1, q2, . . . , q7 were always included. For both female and male individuals, the null hypothesis cannot be rejected, namely that these samples are drawn from the same distribution, as can be seen in Fig. 3.3. Grouped by gender, the distribution of the whole dataset and the sub-dataset for the features *handedness* and *family history of tinnitus complaints* was also compared. For these gender-grouped features, no significant differences between the samples could be revealed. We further compared the baseline characteristics of those individuals that only filled out the baseline characteristics and those that filled out both, baseline and follow-up questionnaires (see Table 3.4). These two groups also show no significant differences in distribution. In addition, the completion for the daily questionnaire differs at a gender-based level and a user-based level. More specifically, most users fill out the daily questionnaire between 1 and 10 times, while others fill it out 100 times or more. The filling-out behavior can be seen in Fig. 3.5. This means that some users are more

represented in the training and testing set than others. However, this does not lead to a different distribution of the baseline characteristics.

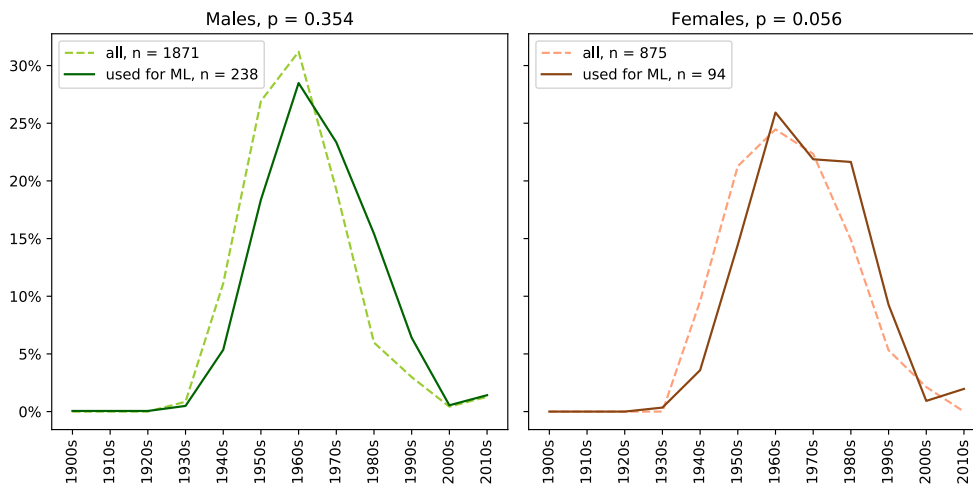


Figure 3.3: The dashed lines denote the age distribution for the all individuals, whereas the solid lines indicate the subset of individuals used for the machine learning calculations. This subset has a size of $m = 11,877$, and contains $238+94$ individual users. For all users, m equals to $80,969$. Note that the high p-values for both groups indicate equality of the age distribution.

Less notably, the gender classification accuracy increases if q_8 (worst symptom) is added. That is due to the fact that there are gender differences in the worst symptom of a tinnitus patient. If a closer look is taken at Fig. 3.4, striking differences can be seen in the distribution of the worst symptom. Women tend to have more difficulties in falling asleep, whereas men tend to suffer relatively more by having difficulties in following a conversation. The authors of [189] revealed similar symptoms of individuals in their work on tinnitus problems. Understanding speech and sleep problems were ranked as the most challenging ones without grouping by gender. The symptom sensitive to environmental noises could be biased by hyperacusis. Individuals with sensitive noise perception would tend to report higher scores here. Since hyperacusis is not assessed in the baseline questionnaire, we cannot consider it. In addition, more factors might bias the discussed symptom (e.g., if one of the parents worked in a noisy factory for a longer period of time, which is not captured by TYT).

When taking a closer look to the correlations of features $q4$ (Mood of user) and $q8_7$ (Depressed because of tinnitus), which is depicted in Fig. 3.6, a negative value can be seen. It is evident why these features should be negatively correlated. An observation with a strong positive correlation appears for the features stressfulness and loudness of

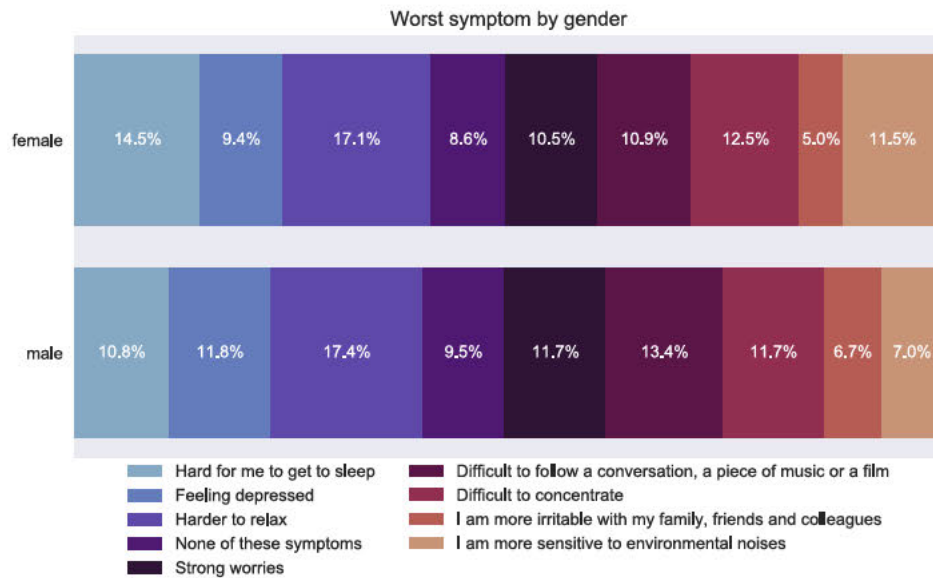


Figure 3.4: Distribution of the worst symptom grouped by gender in a horizontal stacked plot. Each row of the figure adds up to 100%.

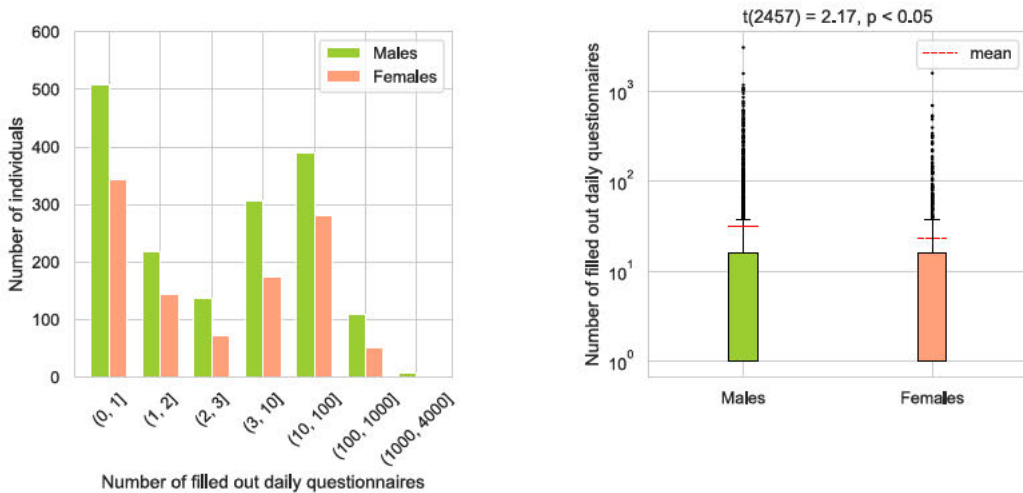


Figure 3.5: Number of filled out daily questionnaires per group (left) and per gender (right). The red-dashed line in the right plot indicates the mean value. Most of the individual users filled out the questionnaire only once. On average, men answered the questionnaire 32 times (+/- 124 std), and women 24 times (+/- 82 std) with $t(2757) = 2.17$ and $p < 0.05$. Notably, there is one male user that filled out the daily questionnaire 3,073 times.

the perceived tinnitus: The louder the tinnitus is, the more stressful it is.

The authors are aware of the trade-off between the depth of a tree within the forest and the standard deviation of the accuracy for a cross-validation set. A higher accuracy could be achieved for a single cross-validation set by increasing the depth of a tree. However, by increasing the depth, a higher variance must be expected between the cross-validation sets, which is an indicator for overfitting of the training set.

For Research Question iii (Which is the most important feature?), the result in the lower-ranked features is ambiguous. For the top three most important features, all three methods rank *strong worries* and *difficulties in following a conversation* firstly and secondly, respectively. For the non-changing questions q1, q2, . . . , q7, however, it is not clear which one could be ranked in the middle or lower for a univariate feature importance. In summary, it can be said that the dynamic question q_8 is rated more important than the non-changing ones.

The results of the presented investigation are both clinically relevant as well as helpful for users of the TYT platform. Regarding clinical relevance, as profound indicators exist that gender differences exist for tinnitus patients [190], TYT can be a valuable alley to learn more about daily fluctuations of tinnitus patients with respect to their gender. As our result show that the answers of the daily questionnaires can predict the gender of TYT users, inversely, the daily answers can be indicators for the symptom differences of men and women. As we further found out that the worst symptom is an important feature, we are in line with other research works beyond the scope of mHealth data [150; 191; 192; 193]. Furthermore, studies that have found gender-related differences in tinnitus patients without using mHealth solutions might particularly benefit from the use of mHealth. For example, in the work presented by [194], it is shown that gender-related differences exist for insomnia. As built-in sensors of smartphones can be used in the context of insomnia [195], mHealth solutions might leverage findings like shown in [194]. Due to the gender-related differences we have found in TYT, it is likely that for other research questions like insomnia mHealth solutions can be helpful as well or even leverage already revealed results. We therefore conclude that in the context of gender-related differences of tinnitus patients, data that were collected with the use of mHealth solutions like TYT are relevant for medical research and clinical practice. Regarding the aspect of helping users with the findings shown here, consider, for example, the work of [191]. One outcome of the latter work describes that anxiety is only associated with bothersome tinnitus in men. Anxiety, in turn, can be easily monitored using a solution like TYT. In this particular case, the gender-related differences can be used to help, for example, men in coping with their anxiety syndrome by learning more about their daily fluctuations (if such fluctuations exist) when using TYT on a daily basis. Inversely, TYT can be used to figure out more variables that are associated with the gender and tinnitus,

which might lead to the development of focused measures that may help to mitigate the tinnitus of men or woman more effectively. To conclude from a tinnitus perspective, TYT has gathered a lot of data and with this data source we were able to reveal that the question on the worst symptom (answered daily) has a high prediction power of the gender of TYT users. Since TYT asks about several worst symptoms, we consider this type of daily questions important. On the other hand, the combination with the other daily questions lead to the final result to predict the gender of TYT users, which we consider as a new outcome of TYT data and research on mHealth in this context.

Overall, the question was investigated whether the answers of male and female tinnitus patients are useful to gain a gender-based differentiation. Therefore, three research questions were investigated: (i) Is it possible to learn a mapping from X to y for the daily tinnitus questionnaire, (ii) which is the most suitable classifier for this task, and (iii) which are the most important features? Four different classifiers of the scikit-learn [173] library from Python were trained to classify the gender of a patient. The most important feature cannot be clearly determined. This result is ambiguous for different feature importance approaches. However, increasing the number of features resulted in a higher classification accuracy. Although the utilization of the possible features showed different results, the gender of the user from a provided daily questionnaire could be revealed with a relevant accuracy. The findings thus might be a valuable basis for the development of more individualized tinnitus treatments, even beyond the scope of TYT.

3.1.4 | Materials and Methods

The study was approved by the Ethics Committee of the University Clinic of Regensburg (ethical approval No. 15-101-0204). All users read and approved the informed consent before participating in the study. The study was carried out in accordance with relevant guidelines and regulations.

The Features For the gender prediction task, two linked data sets were used. The first one, named *Tinnitus Sample Case History Questionnaire (TSCHQ)*, is only provided to an individual *once*, and asks questions like *date of birth, handedness, family history of tinnitus complaints*, the target variable *gender*, and the worst symptom that is related with tinnitus. Baseline characteristics from this questionnaire can be seen in Table 3.4. Note that this table only contains individuals that filled out both, the baseline, and the daily questionnaire. The worst symptom thereby can be one of the following:

- | | |
|--|--|
| <ul style="list-style-type: none"> ■ I am feeling depressed because of the tinnitus. ■ I find it harder to relax because of the tinnitus. ■ I have strong worries because of the tinnitus. ■ Because of the tinnitus it is difficult to follow a conversation, a piece of music or a film. | <ul style="list-style-type: none"> ■ Because of the tinnitus it is hard for me to get to sleep. ■ Because of the tinnitus it is difficult to concentrate. ■ Because of the tinnitus I am more irritable with my family, friends, and colleagues. ■ Because of the tinnitus I am more sensitive to environmental noises. ■ I don't have any of these symptoms. |
|--|--|

The second data set, named *daily questionnaire*, contains daily given answers of a registered individual. This daily questionnaire includes eight questions about the current tinnitus state, i.e., the tinnitus situation and the feelings of the individual *right now*. However, the eighth *dynamic* question depends on the worst symptom of the individual from the TSCHQ questionnaire and asks whether the individual has this specific worst symptom right now or not. If an individual user answered *I don't have any of these symptoms* in the beginning, no question appears in the daily questionnaires. As a consequence, the number of answers for question 8 depends on the number of individuals that have selected this worst symptom in the questionnaire TSCHQ. On the other hand, the number of answers for questions one to seven equals each other. These questions are seen by every individual and are as follows:

- | | |
|--|--|
| <ol style="list-style-type: none"> 1. Did you perceive the tinnitus right now? 2. How loud is the tinnitus right now? 3. How stressful is the tinnitus right now? 4. How is your mood right now? | <ol style="list-style-type: none"> 5. How is your arousal right now? 6. Do you feel stressed right now? 7. How much did you concentrate on the things you are doing right now? 8. <i>This question depends on the worst symptom selected in the questionnaire TSCHQ.</i> |
|--|--|

Depending on the features that are selected for the classification task, the number of examples m depends on the eighth dynamic question.

3.1.4.1 | Data Preparation

The raw data set with the daily answers had the size ($m = 83349$, $n = 19$), where m denotes the number of samples, and n the number of columns. The columns of interest are `individual_id`, `q1`, `q2`, ..., `q7`, `q8_1`, `q8_2`, ..., `q8_8`. In total, the preparation of the data set needed many efforts, namely the following considerations and steps:

The `individual_id` is crucial to merge *TSCHQ* with the daily questionnaire in order to get the gender for a sample of answers. As a consequence, all rows where `individual_id` is NULL were dropped. This affected 1.2% of the samples, i.e., 82,351 samples remained. In the next step, values for `q4`(mood right now) and `q5`(arousal right now) were replaced that have been reported incorrectly from Android devices. For these questions, an individual user can select a position in a self-assessment manikin individual interface feature to represent his or her mood with 9 different steps (i.e., the granularity). However, the Android implementation rounds the values to tenths, which leads to incorrect values. For example, 0.13 has to become 0.125, or 0.88 has to become 0.875.

Missing value treatment As every question is optional, sometimes app users skipped questions. Therefore, the imputation module from the scikit-learn library was used to fill in missing values. In order not to change the data distribution, the data set per individual was calculated. If any of the values for questions 1, 2, ..., 7 was NULL, the missing value treatment was performed. Therefore, the non-null values per column were counted. If there are two or more non-null values, an individual-specific KNN imputation for slider questions with range (0, 1) and Boolean questions [196] was performed. In case an individual user always skipped a specific question, there is no reference how this individual user usually would have answered this question. In such cases, a simple imputation was performed with a median value of the whole data set for slider questions and a most frequent replace for Boolean questions, respectively. An iterative imputation approach was not used as suggested by the authors of [197], because then it would be required to round the estimation of Boolean questions to integer values and fit respective answers to a valid value in $\{0, 0.125, \dots, 1\}$. For the dynamic variable `question8`, missing value treatment does not make sense, as the questions are different. For example, if an individual user has selected *feeling depressed* as a worst symptom, his or her question eight is "Are you feeling depressed right now?". For all the other linked questions, the individual has never seen another dynamic question like "Are you sensitive to environmental noises right now?", as the individual did not report this as the worst symptom. Consequently, these NULL values were left untreated.

3.1. Predicting the Gender of Individuals with Tinnitus based on Daily Life Data of the TrackYourTinnitus mHealth Platform

	meaning	scaling	implementation	count	mean	std
question1	Did you perceive the tinnitus right now?	binary	YesNoSwitch	80969	0.76	0.43
question2	How loud is the tinnitus right now?	continuous	Slider in range (0,1)	80969	0.46	0.3
question3	How stressful is the tinnitus right now?	continuous	Slider in range (0,1)	80969	0.36	0.28
question4	How is your mood right now?	discrete	SAM from 0 to 1 with step size 0.125	80969	0.56	0.21
question5	How is your arousal right now?	discrete	SAM from 0 to 1 with step size 0.125	80969	0.26	0.22
question6	Do you feel stressed right now?	continuous	Slider in range (0,1)	80969	0.28	0.24
question7	How much did you concentrate on the things you are doing right now?	continuous	Slider in range (0,1)	80969	0.58	0.31
question8_0	Because of the tinnitus it is hard for me to get to sleep.	binary	YesNoSwitch	7919	0.35	0.48
question8_1	I am feeling depressed because of the tinnitus.	binary	YesNoSwitch	10361	0.23	0.42
question8_2	I find it harder to relax because of the tinnitus.	binary	YesNoSwitch	13904	0.45	0.5
question8_3	I don't have any of these symptoms.	NULL	NULL	NULL	NULL	NULL
question8_4	I have strong worries because of the tinnitus.	binary	YesNoSwitch	10839	0.27	0.45
question8_5	Because of the tinnitus it is difficult to follow a conversation, a piece of music or a film.	binary	YesNoSwitch	11877	0.33	0.47
question8_6	Because of the tinnitus it is difficult to concentrate.	binary	YesNoSwitch	8220	0.42	0.49
question8_7	Because of the tinnitus I am more irritable with my family, friends and colleagues.	binary	YesNoSwitch	3391	0.32	0.47
question8_8	Because of the tinnitus I am more sensitive to environmental noises.	binary	YesNoSwitch	9179	0.09	0.29
gender	0 = Male, 1 = Female	binary	Single Choice	80969	0.26	0.44

Table 3.3: Description of the data frame used for the machine learning approaches. Note that the count for the questions 8_0, 8_1, . . . , q_8 is dependent on the number of individuals that selected this answer in the baseline questionnaire. If an individual selected *I don't have any of these symptoms*, no follow-up question appeared, so that these values are NULL. SAM = Self-Assessment Manikin [198].

Calculation of the Correlation Matrix The values of Fig. 3.6 were calculated using three different methods depending on the scaling of the features. Note that it is not possible to calculate the correlations of the q8 questions to each other as they are pairwise disjoint. If both features are continuous, the Pearson correlation has been used [199]. If one feature is either discrete or binary and the other is continuous, the Pointbiserial correlation was calculated [200]. Finally, if both features are discrete or binary, the Corrected Cramer's V correlation has been calculated [201]. Further note that Cramer's V correlation is defined for a range of (0,1), whereas Pearson and Pointbiserial for a range of (-1,1).

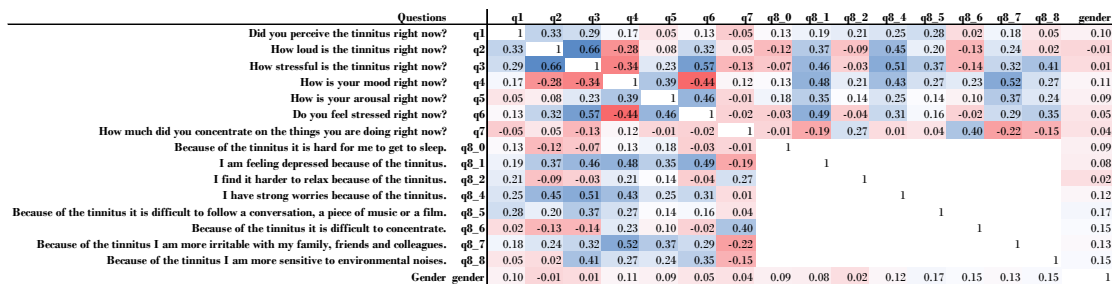


Figure 3.6: Heatmap for feature-gender cross-correlations. The last column (resp. the last row) shows the correlation of the whole data set (without equal splits for male and female individuals) with the target gender. Depending on the feature scaling, different correlation approaches (Cramer's V, Pointbiserial and Pearson) have been used. The matrix reveals strong positive correlations between stressfulness and loudness of the tinnitus or negative correlations between mood and stressfulness of an individual user. The heatmap was formatted using MS Excel 365. Correlation metrics were calculated using SciPy 1.5.0 within a Python 3.7 environment.

Univariate Feature Classification For this classification task, a random forest classifier was used as proposed by the authors of [46]. In order not to get a biased estimation of the feature importance, a grouped data set per feature was calculated. As can be seen in Table 3.3, the number of examples n varies per feature. Therefore, the feature was taken with the smallest training examples (q8_7), and randomly 50% men and 50% women from the target gender were selected. In the next step, X was defined as the feature space of shape (m, n) , with m = number of examples, and $n = 1$, as only one feature was used. Then, a Random Forest classifier from scikit-learn was instantiated, including 80% of randomly chosen examples, which denotes the training set. Next, the accuracy on the remaining 20% of the examples was calculated, which denotes the testing set. Note that there is no development set for this subtask, as hyper-parameter tuning is not performed initially. For each feature, this procedure was repeated 10 times and the mean of those 10 accuracies were determined. The features q8_4, q8_5 (worries, difficulties in following a conversation) and q8_6 (difficulties in concentration) reach accuracy values greater than 0.58, which is significantly better than random guessing. Consequently, these features are ranked top three.

Comparison Comparing the results of the three feature importance approaches, the result for the top two features is unambiguous. However, the correlation approach ranks *sensitivity on environmental noises* on a third place, whereas the permutation and random forest approach *difficulties in concentration* have different results on this rank place.

3.1.4.2 | Supervised Machine Learning Application

Feature Selection After determining which variables were more and which less important for a univariate approach, the best set of features (multivariate approach) had to be identified in order to find a mapping from X to y , where X is a subset of all features and y is a binary gender prediction with male and female individuals. However, an arbitrary combination of features is only possible within the feature set of $\{q1, q2, \dots, q7\}$. Only one out of the features from question 8 can be added optionally. This constraint leads to 1,143 valid subsets of the data set. In order to get the best feature list, every single combination of valid subsets to an 80-20 training-testing-split of the data set was applied, before storing its accuracy and the corresponding feature list to a Python dictionary. Given a Random Forest classifier, it can be simply said that a feature list is superior to another if its accuracy on average in the testing set is higher. Without any of the dynamic questions from $\{q8_0, q8_1, \dots, q8_8\}$, the best set contains the features $\{q2, q3, \dots, q7\}$. Note that $q1$ is not included. This set leads to an accuracy of **72.7%**, with a testing size

of $n = 8276$. If one of the q8-questions is added to the feature set, the most promising combination contains {q1, q2, ..., q7, q8_5}, with an accuracy of **81.7%** on average, and a test size of $n = 1702$.

Classifier Comparison This section covers aspects to address Research Question *ii*: Which machine learning model is most suitable for predicting the gender of an individual user and has a high prediction power? More specifically, four supervised machine learning classifiers were investigated: A Support Vector Machine [202], a Multilayer Perceptron Neural Network (MLP) [203], a Decision Tree [204] and a Random Forest [46]. With the same testing size from the previous section of $n = 1702$, the following results were obtained. The Decision Tree reached the lowest accuracy with 79%, followed by the Support Vector Machine with 80%, and the Multilayer Perceptron with 81%. The Random Forest classifier reached 86% in accuracy in the best cross-validation set. The ROC curve in Fig. 3.7 affirms the superiority of the Random Forest classifier for this specific classification. The Support Vector Machine and the Multilayer Perceptron have a very similar performance. The Decision Tree contains only pure subsets in its final leaves, which leads to a triangled ROC curve and in this case, eventually meaning the lowest performance.

Hyper-parameter Set-up In a first approach, the four classifiers have been used mainly with a default set from the Python scikit-learn library [173]. Then, several hyper-parameters were slightly adjusted, i.e., the number of neurons per layer for the Multilayer Perceptron Regressor, and the splitter criterion for the Decision Tree classifier. The details of the hyper-parameters can be seen in Listing 3.2.

```

1 SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
2     decision_function_shape='ovr', degree=3, gamma='scale', kernel='rbf',
3     max_iter=-1, probability=False, random_state=1994, shrinking=True,
4     tol=0.001, verbose=False)
5
6 DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
7     max_depth=None, max_features=None, max_leaf_nodes=None,
8     min_impurity_decrease=0.0, min_impurity_split=None,
9     min_samples_leaf=1, min_samples_split=2,
10    min_weight_fraction_leaf=0.0, presort='deprecated',
11    random_state=1994, splitter='best')
12
13 MLPClassifier(activation='tanh', alpha=0.0001, batch_size='auto', beta_1=0.9,
14    beta_2=0.999, early_stopping=False, epsilon=1e-08,
15    hidden_layer_sizes=(8, 16, 32, 2), learning_rate='adaptive',
16    learning_rate_init=0.001, max_fun=15000, max_iter=500,
17    momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True,
18    power_t=0.5, random_state=1994, shuffle=True, solver='adam',

```



```

19         tol=0.0001, validation_fraction=0.1, verbose=False,
20         warm_start=False)
21
22 RandomForestClassifier(bootstrap=False, ccp_alpha=0.0, class_weight=None,
23         criterion='entropy', max_depth=50, max_features='auto',
24         max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.0,
25         min_impurity_split=None, min_samples_leaf=1, min_samples_split=2,
26         min_weight_fraction_leaf=0.0, n_estimators=1000, n_jobs=None,
27         oob_score=False, random_state=1994, verbose=0, warm_start=True)
    
```

Listing 3.1: Hyperparameter set-up for the used classifiers

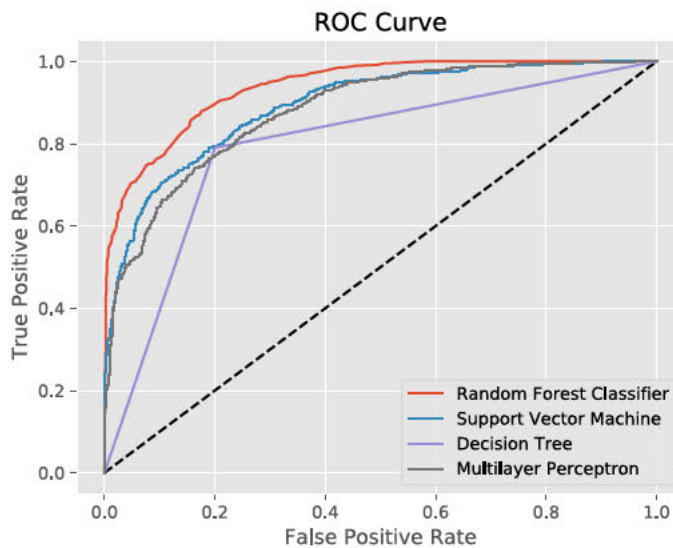


Figure 3.7: ROC curve for compared classifiers. As the decision tree contains only pure subsets, the class probabilities are either 0 or 1. This leads to a triangled ROC curve.

According to the classifier’s accuracy, the Random Forest classifier seems to be most suitable for this task, which was used to answer Research Question *ii*.

Characteristic						
	n	Age (<i>std</i>)	Right-handed	Left-handed	Both sides	Existing family history of tinnitus complaints
Male	1871	49.23 (14.40)	1345 (71%)	282 (13%)	244 (16%)	426 (23%)
Female	875	46.04 (14.72)	650 (75%)	126 (11%)	99 (14%)	235 (27%)
Total	2746	48.71 (14.89)	1994 (73%)	408 (12%)	343 (15%)	661 (24%)

Table 3.4: Baseline characteristics of the Tinnitus Sample Case History Questionnaire (TSCHQ) for all individuals that filled out at least one follow-up questionnaire. Individual users that registered for this study, but did not fill out at least one follow-up questionnaire, are not considered in this table.

Supplementary Information

The Python code to replicate the Machine Learning classifiers, figures and tables is available on [Github](#).

3.2 | Prediction of Tinnitus Perception based on Daily Life mHealth Data using Country Origin and Season

- **Authors** | Allgaier, Johannes; Schlee, Winfried; Probst, Thomas; and Pryss, Rüdiger
- **Published in** | Journal of Clinical Medicine, 11(15), 4270, 2022.
- **Available at** | <https://pubmed.ncbi.nlm.nih.gov/35893370/>

Abstract

Tinnitus is a phantom auditory perception without external sound stimulations. The chronic perception of tinnitus can severely impact the quality of life. As tinnitus is characterized by a heterogeneity of the patient's symptoms, researchers often use a multi-modal data fusion approach to reveal new insights. However, differences across countries and seasons based on mobile health data have been not presented so far. Therefore, data of the **TrackYourTinnitus** (TYT) mHealth platform were investigated to see whether season-related differences in the symptom profiles of TYT users exist. In addition, differences based on the country origin were investigated. The conducted analyses address three major research questions. First, it was analyzed whether the momentary tinnitus can be related to the season or country origin. We used a gradient boosting machine (GBM) to binarily classify the momentary tinnitus on the assessment level with an accuracy of 94.03 %. Second, another GBM was trained to regress the tinnitus loudness on a scale from 0 to 100. On the daily assessment level, the tinnitus loudness can be regressed with a mean absolute error rate of 7.9 %-points. Both results indicate differences in tinnitus of TYT users with respect to the season and country origin. Third, country- and season-specific differences were analyzed. It could be revealed that tinnitus varies with the temperature in certain countries. The considered perspectives, in turn, have been derived through the inspection of the TYT data set and its possibilities. The presented results show that the season and the country origin seem to be valuable features when being combined with longitudinal mHealth data on the daily assessment level.

3.2.1 | Introduction

Tinnitus is widely known as a long-term noise in the ears, which is described by patients through heterogeneous sound manifestations [138]. Economically, tinnitus induces a

high burden, as about 10 - 15% of the worldwide population [139; 140] is affected by this chronic disorder. 2.4% of these affected patients severely suffer from tinnitus day by day [141], while one to two percent experience a reduction in their quality of life due to tinnitus, including insomnia, anxiety, hearing difficulties, or depression [205; 206; 207]. At present, no general treatment exists, which is able to effectively reduce *tinnitus loudness* and related fluctuations. Consequently, many patients are confronted with a complex healthcare situation, which often reduces the quality of life significantly. The mentioned heterogeneity of tinnitus symptoms also complicates the development of new and more general treatment methods [143; 144]. However, on an individual basis, tinnitus can be reduced, for example, by the use of cognitive behavioral therapies [142].

Various efforts are constantly made to learn more about the heterogeneity of symptom profiles of tinnitus patients. However, data sources are often missing to investigate aspects with respect to this heterogeneity of symptom profiles that seem to be interesting. As the proliferation of smartphones has led to powerful mobile health solutions (denoted as mHealth solutions) that are able to establish data sources with opportunities to better deal with differences of symptom profiles, in this paper, such mHealth data source is investigated for tinnitus patients. Although respective investigations have gained attention recently, many opportunities are still not utilized. For example, a comparison of mHealth data of tinnitus users across countries does not exist to the best of our knowledge. In addition, detailed insights based on season differences are also less considered in the context of collected mHealth tinnitus data so far. Therefore, these two questions on differences across seasons and the country origin have been selected for further investigations on symptom profiles of tinnitus patients using a mHealth platform. In the context of the mentioned differences, only little research has been presented. In addition, these presented works are all beyond the scope of mHealth. There is one study on seasonal changes in tinnitus symptomatology, which concludes that searches for tinnitus aspects are higher in winter than in summer in some countries [208]. Another work suggests an association of depression, a common comorbidity of tinnitus, and season. It provides Internet-based evidence for the epidemiology of seasonal depression. The results suggest that Internet searches for depression by people at higher latitudes are more affected by seasonal changes, while this phenomenon is faded out in tropical areas [209]. However, already more than 70 years ago, it was clinically observed that tinnitus increases during the winter months [210; 211]. Seasonal affective disorders (SAD), in turn, were studied by the authors of [212]. They conclude that SAD are present when a symptom occurs during the winter months and disappear completely in summer.

When aiming at mHealth solutions to investigate these differences, at first, the type of collected data must be taken into account as mHealth solutions can be based on different

methods, strategies and concepts. In this work, Ecological Momentary Assessments (EMAs) are the basis for the investigations as they are particularly appropriate for the investigations at question [2]. However, EMA only defines the strategy how participants of a study (usually, longitudinal studies) will be questioned. Three aspects are the main pillars of the EMA strategy: EMAs must be carried out in real life (opposed to a clinical environment) and at arbitrary points in time (to capture the moment of a participant). Third, a concrete measurement (e.g., though a questionnaire) must be accomplished. If EMAs are now performed through the boundaries of a year and across countries, a data source can be established through such measurements that enables a powerful basis to investigate country- and season-specific differences. Recall that EMA only defines the strategy. In the context of mHealth, digital phenotyping techniques[213] express an important trend to use smartphones to practically enable Ecological Momentary Assessments (EMAs). Digital phenotyping quantifies the human phenotype in a moment-to-moment fashion using active and passive data from mobile devices. As smartphones are present in daily life of almost anyone, the performance of EMAs through smartphones can effectively capture the daily life of users over time. Respective evaluations based on mHealth data, in turn, have been recognized as potential alleys for a better support of patients [214]. mHealth apps, in turn, are the major instrument to operationalize digital phenotyping and EMAs. Many mHealth apps have been presented in this context [10; 154; 161]. Although valuable data sources have been established by the use of digital phenotyping, mHealth data comes also with drawbacks [155], which must be considered carefully. For example, in EMA settings, in which users fill out several questionnaires each day over a longer period of time, it must be ensured that the data was provided in a meaningful way. To get a better impression regarding the meaningfulness, the following example shows emerging challenges through EMA. If users have to fill out a lot of questionnaires through EMAs more than once a day, then they could tend to fill out only to accomplish the task itself, without providing the actual momentary situation. In the context of tinnitus, the TrackYourTinnitus platform (TYT), which is based on mobile crowdsensing techniques [158] as well as EMAs [30], puts digital phenotyping into practice. Crowdsensing, in turn, connects a group of people, who have mobile devices with sensing and computing capabilities, collectively sharing data, and extracting information to measure, map, analyze, and estimate any processes of common interest. TYT was initially developed to investigate questions about the aforementioned heterogeneity of symptom profiles of tinnitus patients [159; 160; 161]. The procedure how users are walking through TYT is described in [172]. In essence, users register to the platform (website or mobile apps), then they have to fill out three baseline questionnaires asking about demographic data and tinnitus characteristics. The users have to fill out these

questionnaires before they are able to start with the EMA procedure. The latter is applied through two native apps, which are available for [iOS](#) and [Android](#) in the official app stores. The EMA procedure consists of a daily questionnaire with up to eight questions. This questionnaire is applied using two strategies. The first one is based on the idea that users can fill out the questionnaire whenever they want. The second strategy is based on notifications. Up to 12 random notifications or a fixed schema are used (can be chosen by users, which schema they prefer) to remind the users to fill out the EMA questionnaire. The mainly used schema are the random notifications [163]. As this selection follows the idea of in situ measurements in the sense of digital phenotyping, many investigations and analyses become possible. Of further importance, until today, this setting has motivated over 8000 users from all parts of the world to provide more than 100,000 questionnaires. The use of mHealth in this context, apart from TYT, has been proposed by many other mHealth projects [142; 164; 165], which indicates that strategies like EMA or digital phenotyping are promising in the context of tinnitus research. The mentioned investigations on differences across seasons and the country origin have been identified to be possible on the TYT data source [163]. For the concrete analyzes, we have decided to work on the following three major research questions (RQ):

- RQ1: Can the *momentary tinnitus* (Question 1 of the daily EMA questionnaire; yes/no answer options) of TYT users be predicted (i.e., a binary classifier be trained using machine learning based classifiers) using the features country, season, age, and sex as well as the daily EMA questionnaire and its questions on mood, arousal, stress, concentration, and the worst symptom perception?
- RQ2: Can the reported *tinnitus loudness* of TYT users (Question 2 of the daily EMA questionnaire; slider question) be predicted (i.e., a regressor be trained using machine learning based classifiers) based on the same features like for RQ1?
- RQ3: Based on inferential statistics, are we able to reveal country- and season-specific differences for the reported *momentary tinnitus* based on the daily EMA questionnaires of TYT users?

Regarding RQ1 and RQ2, we will present results from two machine learning analysis. As TYT was able to gather more than 100,000 EMA questionnaires since 2013, which are comprised of many dimensions, we decided to answer RQ1 and RQ2 based on machine learning algorithms. As we already revealed interesting results on TYT EMA-data based on machine learning [167] as well as the use of machine learning has been generally recognized in the context of mHealth data in the last years with much attention and valuable results [166; 168; 169; 170], the following paper links up with these findings.

Regarding RQ3, we will present descriptive statistics about the identified country- and season-specific differences. We have detailed the research question into four sub-questions due to the following reason: Based on the two main goals to investigate country- and season-specific, which represent the two categories of differences, we were able to derive further promising questions. RQ3₃ is a combined perspective of the country and the season, while RQ3₄ is inspired by medical experts. The following list presents the four sub-questions:

- i RQ3₁: Are there country-specific differences for the *momentary tinnitus*?
- ii RQ3₂: Are there season-specific differences for the *momentary tinnitus*?
- iii RQ3₃: In the light of a combination of country- and season-specific differences, the question arose, whether the *momentary tinnitus* varies within the year and across countries.
- iv RQ3₄: Another question arose, whether country- and season-specific differences of the reported worst symptom can be identified.

Three additional notes are important regarding RQ3₁-RQ3₄. First, the last sub-question was set up due to the involved medical experts as severe symptoms play an important role in the context of tinnitus research. As TYT asks about nine possible worst symptoms, we investigated how the worst symptom differs across countries and seasons. As the combined perspective taken for RQ3₃ was useful, this combined perspective was also accomplished for RQ3₄. Second, in the context of season-specific differences, we added an additional dimension, the temperature course throughout the year, which is inspired by the results of [209].

Finally, for the prediction tasks in RQ1 and RQ2, we excluded features of TYT that are highly correlated with the target, such as *tinnitus loudness*, *tinnitus stress*, and *momentary tinnitus*. However, we included features that are known to be correlated with tinnitus, such as sex and age [215].

3.2.2 | Materials and Methods

The study was approved by the Ethics Committee of the University Clinic of Regensburg (ethical approval No. 15-101-0204). All users read and approved the informed consent before participating in the study. The study was carried out in accordance with relevant guidelines and regulations.

The questionnaires For the tinnitus prediction task, three linked data sets were used. The first (1) one refers to the baseline questionnaire named *Tinnitus Sample Case History Questionnaire (TSCHQ)*. The second one (2) is the daily questionnaire and asks for information about the user's current sense of well-being. The third data set (3) contains information on the temperatures of a country in the annual cycle.

The first (1) *TSCHQ* questionnaire is completed by each TYT user *once* when starting the app for the first time. In this questionnaire, demographic data as well as data about the individual course of the tinnitus are collected, such as the onset of the tinnitus, or the worst symptom that is related to tinnitus.

When logging in into the TYT platform, users are asked for their worst tinnitus symptom. This symptom can be one of the following.

- I am feeling depressed because of the tinnitus.
- I find it harder to relax because of the tinnitus.
- I have strong worries because of the tinnitus.
- Because of the tinnitus it is difficult to follow a conversation, a piece of music or a film.
- Because of the tinnitus it is hard for me to get to sleep.
- Because of the tinnitus it is difficult to concentrate.
- Because of the tinnitus I am more irritable with my family, friends and colleagues.
- Because of the tinnitus I am more sensitive to environmental noises.
- I don't have any of these symptoms.

As we also record fill-in dates of answers to this questionnaire, and the country of the user, we can link the worst symptom to both the season and country. To assign the fill-in date to a season, we used the astronomical seasons as a guide. More specifically, spring starts on March 21st, summer in June 21st, autumn in September 23rd, and winter in December 21st. For countries of the southern hemisphere, the seasons are opposite, i.e., spring becomes autumn, summer becomes winter. etc. 3.2 % of the collected data comes from countries in the southern hemisphere. The correction of the seasons concerns only the analysis for the worst symptom. For the machine learning part (RQ1, RQ2), countries in the southern hemisphere were not involved due to the insufficient number of completed questionnaires.

The second (2) data set refers to the *daily questionnaire*. It includes eight questions about the current tinnitus state, i.e., the tinnitus situation and the feelings of the individual

right now. However, the eighth *dynamic* question depends on the worst symptom of the individual from the TSCHQ questionnaire and asks whether the individual has this specific worst symptom right now or not. If an individual answered *I don't have any of these symptoms* in the beginning, no eighth question appears in the daily questionnaire. Consequently, the amount of data for question 8 depends on the number of individuals that have selected this worst symptom in questionnaire TSCHQ. On the other hand, the number of answers for questions one to seven equals each other. These questions are seen by every individual in the same way and are as follows:

1. Did you perceive the tinnitus right now?
2. How loud is the tinnitus right now?
3. How stressful is the tinnitus right now?
4. How is your mood right now?
5. How is your arousal right now?
6. Do you feel stressed right now?
7. How much did you concentrate on the things you are doing right now?
8. *This question depends on the worst symptom selected in the questionnaire TSCHQ.*

Depending on the features that are selected for the classification task, the number of examples m depends on the dynamic question eight. The questions for *mood* and *arousal* are questions using a self-assessment scale (SAM) [172], with 9 possible values. Depending on a user's operating system, the answer is stored with different accuracy. Therefore, rounding errors can occur in the hundredths range on Android phones. We neglected these rounding errors in pre-processing considering the amount of 18 other features (*countries, seasons, sex, age, mood, arousal, stress, concentration, worst symptom perception*. Note that *countries* and *seasons* are categorical and thus one-hot encoded features.), as described in Table 3.5.

The third (3) dataset contains information about the temperature in the country per season. The dataset was crawled from Wikipedia and is originally a list of cities with their average monthly temperatures. The respective country is noted for the cities, which means that many countries are represented by several cities. In this case, the data was grouped by country and averaged again. The weather data in the cities themselves are taken from various weather services in the respective countries, which sometimes results in temporal differences in the data, which we consider negligible, however, due to the slow climate change. The temperature dataset can directly be found [here](#). The mapping of the country names with the iso2 country codes (i.e., *Federal Republic of Germany: DE*) was done using [this](#) list.

3.2.2.1 | Data preprocessing

The raw data comes from three .csv files, which, in turn, are extractions from the TYT database [163]. The first file is a data frame containing meta information from all registered users (number of users = 8685 by Feb. 2021). This meta data includes, among others, the country, nationality, and mobile platform. The second file is the baseline questionnaire and contains 3700 users that filled out the initial questionnaire. The daily questionnaire is the last file with 3044 users that answered 98,074 daily questionnaires. We can see from this, that of the registered users, about one in three completes the daily questionnaire at least once.

The *user_id* is mandatory to merge the three data sets. As a consequence, all rows where *user_id* equals *NULL*, we dropped that row. We further removed the 25 test-users with known user IDs to reduce bias and noise in the data. The remaining merged data frame had 97,742 rows and 65 columns. This data frame has been used for the statistical analyses provided in the results section.

Machine Learning Preprocessing For the machine learning task, a further preprocessing was required. Gradient boosting machines can only handle numerical data with no missing values. We therefore dropped rows that contained missing values, which affected about 24 % of the data. We then needed to convert categorical features into numbers. As decision trees split data in binary groups, we used the *pandas.get_dummies()* method to convert the countries and seasons into several columns. The column name is then the category. A 1 indicates that this category applies, i.e., *autumn = 1*, which, in turn, means that the other seasons must be zero. In order not to increase the number of columns unnecessarily, we used the *drop_first = True* keyword argument. This means, we get k-1 dummies out of k categorical levels by removing the first level. The last step considered the imbalanced distribution of the target variable *tinnitus occurrence*. About 79 % of the assessments state *yes*. Any naive machine learning classifier would therefore simply always predict *yes*, regardless of the input of features and would still get 79 % accuracy on average. Using the F1 accuracy score, the performance can be measured better, but the classifier would still be overfitted on positive examples. We therefore bootstrapped negative examples with replacement until we had a balanced dataset. The final dataset had 118,054 samples with 22 features each.

Estimation of feature importances The values of Table 3.6 were calculated using three different methods, the Gini importance, the permutation importance, and the correlation metric. Depending on the feature scaling, two different correlation metrics have been

variable name	variable meaning	mean	std	scaling
AT	Austria	0.02	0.13	
CA	Canada	0.03	0.16	
CH	Switzerland	0.08	0.27	
DE	Germany	0.62	0.49	
GB	Great Britain	0.05	0.21	
IT	Italy	0.01	0.10	
NL	Netherlands	0.07	0.25	
NO	Norway	0.02	0.13	binary
RU	Russia	0.02	0.14	
US	United States	0.09	0.29	
spring		0.26	0.44	
summer		0.24	0.43	
autumn	season	0.25	0.43	
winter		0.25	0.44	
Male	Sex	0.74	0.44	
age	Age in years	49.71	12.98	integer
question4	How is your mood right now?	0.58	0.20	
question5	How is your arousal right now?	0.25	0.22	SAM from 0 to 1 with stepsize 0.125
question6	Do you feel stressed right now?	0.26	0.23	
question7	How much did you concentrate on the things you are doing right now?	0.59	0.31	Slider in range (0, 1)
question1 (target RQ1)	Did you perceive the tinnitus right now?	0.50	0.50	binary
question2 (target RQ2)	How loud is the tinnitus right now?	0.47	0.30	Slider in range (0, 1)

Table 3.5: Overview of the features and the targets used to train the gradient boosting machines for RQ1 and RQ2. Most of the features are binary, *age* has the highest cardinality. The whole dataset had the shape (118054, 22) after re-balancing for the target *momentary tinnitus*. For the ML feature, the average *age* is higher as some users completed the questionnaire over several years and *age* was calculated at the time of completing the daily questionnaire.

applied. If the input feature was categorical, Corrected Cramer’s V [201] was applied. If it was continuous, the Point Biserial method [200] was used. Cramer’s V is defined in range (0, 1), whereas the Point Biserial correlation is defined in range (-1, 1). Nevertheless, to be able to order the results *within* the column, we took the absolute value from the Point Biserial result. Although all results are in percentages, it is not possible to compare them line by line. This is due to the different units of measurement. Therefore, we have created the ranking. For the Gini and the permutation importances, both methods are used using the trained gradient boosting machine. The Gini importance is an impurity-based method. The higher it is, the more important the feature is. Notably, within this column,

all values add up to 100 %. The importance of a feature is calculated as the reduction of the impurity caused by this feature. For the permutation importance, the percentage values are an estimate for the increase of the error rate on average, if that feature would have been replaced by a random feature. That means, if the variable *gender* would be replaced with a random variable, the error would increase by 6.43 %-points. That column does not necessarily add up to 100 %.

3.2.2.2 | Gradient Boosting Machines for classification of *momentary tinnitus* and regression of *tinnitus loudness*

Why did we choose the Gradient Boosting Machine? It is a tree-based Machine Learning algorithm and related to Random Forests. Machine Learning contests on the Kaggle platform have recently shown that this algorithm is superior to most state-of-the-art Deep Learning methods when it comes to tabular data, such as house pricing prediction problems. Both, Random Forests and Gradient Boosting Machines use several trees to predict the outcome. However, one of the main differences between those two algorithms is the *time aspect*. That is, the Gradient Boosting algorithm learns from previous misclassified samples by putting more weight on those. Furthermore, it does not easily tend to overfitting like decision trees do.

We used the Python implementation from scikit-learn [173] to apply the Gradient Boosting machine to the dataset. We then defined the 20 features (10 countries, 4 seasons, sex, age, mood, arousal, stress, concentration level) and the targets (*momentary tinnitus*, *tinnitus loudness*). The whole dataset was divided into three sets: Training, development, and testing. Training plus development got 70 % of the data, testing 30 %. To avoid a selection bias within the classification problem, we stratified on *y*. Setting a *random_state* (also known as seed) ensured that the results are reproducible. For the tuning of the hyperparameters, we used a gridsearch approach. Within that, we varied the *learning_rate*, the *max_depth* of each tree, the sizes of the *subsamples*, the minimum number of samples per leaf, and the fraction of randomly chosen features per tree. 1,280 combinations of the hyperparameters have been evaluated systematically, the final chosen setup can be seen in Listing 3.2 for the classifier and Listing 3.3 for the regressor, respectively. Each combination was cross-validated within the training set using a 5-fold split. This means that the 70 % of the training data was further divided into 5 folds. Four of each were used for training and one for validation.

For the classification task, the mean test accuracy score on validation was 91.1 % (std = .002). On the test dataset, an even higher accuracy of **94.03 %** was achieved.

When leaving out the features *sex* and *age*, the mean test score dropped to 88.9 % using

the same hyperparameters. Using only the binary features *seasons* and *countries* leads to a decrease of the accuracy on the test set down to 58 %. This is caused by the low dimensional feature space.

For the regression task, a mean absolute error of 8.1 % was achieved on the validation set (std = .0006) and a 7.9 % error on the test set.

```

1 # Gridsearch setup
2 params_gb = {'learning_rate': [
3     0.1, 0.2, 0.3, 0.5, 1],
4     'max_depth': [3, 4, 5, 10],
5     'verbose': [1],
6     'random_state' : [42],
7     'subsample': [0.25, 0.5, 0.75, 1],
8     'min_samples_leaf': [1, 2, 3, 10],
9     'max_features': [0.25, .5, .75, 1]
10 }
11 # Chosen hyperparameters
12 GradientBoostingClassifier(loss='deviance', learning_rate=0.5,
13     n_estimators=100, subsample=1.0, criterion='friedman_mse',
14     min_samples_split=2, min_samples_leaf=1,
15     min_weight_fraction_leaf=0.0, max_depth=10,
16     min_impurity_decrease=0.0, min_impurity_split=None, init=None,
17     random_state=42, max_features=0.5, verbose=0,
18     max_leaf_nodes=None, warm_start=False,
19     validation_fraction=0.1, n_iter_no_change=None, tol=0.0001,
20     ccp_alpha=0.0)

```

Listing 3.2: Hyperparameter set up for the Gradient Boosting Classifier

```

1 # Gridsearch setup
2 params_gb = {'learning_rate': [
3     0.1, 0.2, 0.3, 0.5, 1],
4     'max_depth': [3, 4, 5, 10],
5     'max_features': [0.25, .5, .75, 1],
6     'random_state' : [42],
7     'subsample': [0.25, 0.5, 0.75, 1]
8 }
9
10 # Chosen hyperparameters
11 GradientBoostingRegressor(
12     learning_rate=0.5,
13     max_depth=10,
14     max_features=0.75,
15     random_state=42,
16     subsample=1,
17     verbose=1)

```

Listing 3.3: Hyperparameter set up for the Gradient Boosting Regressor

3.2.3 | Results

In this section, the results for the research questions are presented subsequently. At first, we focus on the first question of the daily TYT questionnaire (*Did you perceive the tinnitus right now?*). We refer to this question as the *momentary tinnitus* in the following. Second, we consider the *tinnitus loudness* (*How loud is your tinnitus right now?*) and refer to this question as *tinnitus loudness*. Third, we analyze these two targets *momentary tinnitus*

and *tinnitus loudness* in a global context by relating them to the country, season, and temperature.

Features for RQ1 and RQ2 We used four different groups of features. The first group of features are dummy features indicating whether an individual comes from that country or not. As 111 countries would lead to an unnecessary increase in the size of the features, we only took those 10 countries with the most filled out daily questionnaires. These countries are ['DE', 'US', 'NL', 'CH', 'GB', 'CA', 'RU', 'AT', 'IT', 'NO']. The second group of features are the four *seasons*, which are also coded as dummy features. The third group contains *age* and *sex*. Note that we did not include the questions *tinnitus loudness*, *momentary tinnitus* and *tinnitus stress level* as features as they are highly correlated with the respective target.

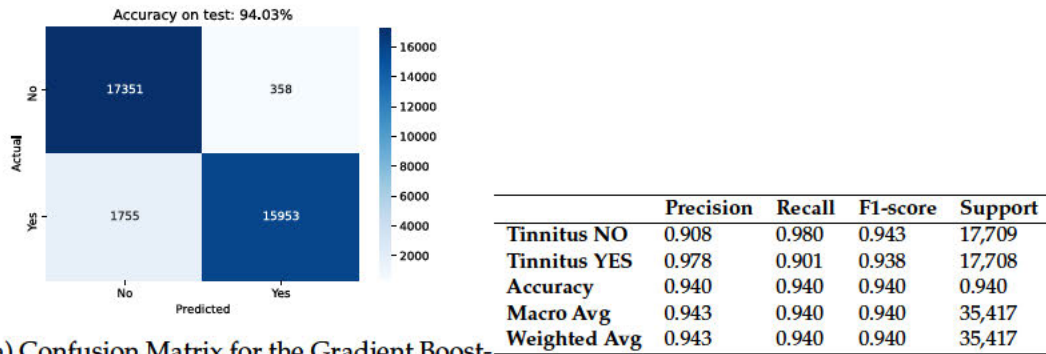
The age is calculated from the date of the completed daily questionnaire and the date of birth. Sex contains two unique values, male and female. The last group of features is a subset of questions of the daily questionnaire. This subset contains information about the momentary mood, arousal, stress level, and concentration. This results in a data frame with 20 features, 1 binary target, and 74,360 samples from 2,179 users.

3.2.3.1 | RQ1: Is the *momentary tinnitus* of TYT users predictable using the features *country, season, age, sex, and from the daily EMA questionnaire, mood, arousal, stress, concentration, and worst symptom perception*?

Data preparation From previous research works, we already knew that the dataset is imbalanced regarding the target. This means that about 75,000 answers are *tinnitus = yes*, but only 20,000 *tinnitus = no*. A classifier that has guessed randomly the outcome would get 50 % accuracy on average, a *naive* classifier would simply always predict *Tinnitus=yes* and would get 78.95 % accuracy on average. We therefore draw randomly 54,566 times a sample from the *Tinnitus=no* group, add it to the data frame, and finally shuffle the samples. This forces each naive classifier to an accuracy down to 50 %. This, in turn, means that any improvement in the accuracy can be attributed to the learning of the classifier.

Machine Learning The machine learning task at hand is a binary classification task. We wanted to know whether it is possible to predict the occurrence of tinnitus for an individual of the TYT platform. We used a Gradient Boosting Machine [89], which builds an additive model and learns subsequently from prior classification trees. We further divided the data into three sets: Two for cross-validation (the training and the validation

sets), and one for the final testing. For cross-validation, we used 70 % for the testing, and 30% for the validation. We stratified on y while splitting in order to retain the 50-50 distribution of the binary target. After a hyper-parameter tuning using gridsearch, we got a **final accuracy of 94.03 %** in the testing set. Details are provided in Fig. 3.8.



(a) Confusion Matrix for the Gradient Boosting classifier. Although there is a little tendency on false negatives, the overall accuracy of 94.03 % on the test set is significantly better than random guessing. (b) Classification report for the Gradient Boosting Classifier. The tendency to predict *No* rather than *Yes* leads to a larger F1 score for Tinnitus=*No* and a larger recall for Tinnitus=*Yes*.

Figure 3.8: Confusion matrix and classification report for the Gradient Boosting Machine used to predict whether an individual has *momentary tinnitus* or not.

Feature Importance To find out which of the variables have a high impact on tinnitus prediction, we looked at the feature importance of the Gradient Boosting machine. In order to determine the feature importance more accurately, we have investigated three methods for this. The first one is called *Gini importance*, the second one is *permutation importance*, while the last one is the *correlation*. These three methods measure the feature importance in different units, which makes it impossible to compare importances between methods. However, to make the results comparable, we have created an importance ranking. The lower (i.e., greener) the ranking number is, the more important the feature for the model to predict the target is. In Table 3.6, we separated the features into four groups: Countries (with ISO2 country codes), seasons (spring, summer, autumn, winter), demographics (age, sex), and daily questions (mood, arousal, stress, concentration). We further calculated the feature importances for each model (classifier and regressor) separately. To classify the *momentary tinnitus*, demographic features are most important with an average rank of 4.5. The average rank is calculated as the mean of all ranks (Gini, permutation, correlation), for the features belonging to that group. To regress the *tinnitus loudness*, the daily questions are most important with an average rank of 4.16). For both models, *age* is the most important feature (average rank = 2), as it has a high cardinality.

Conversely, the countries have a lower importance (average rank = 13.6), since they have only a low cardinality with low variance.

Feature	Did you perceive the tinnitus right now? - Classification						How loud is the tinnitus right now? - Regression					
	Gini	Permutation	Correlation	Gini Rank	Perm. Rank	Corr. Rank	Gini	Permutation	Correlation	Gini Rank	Perm. Rank	Corr. Rank
AT	0.4%	0.2%	2.9%	20	20	14	0.4%	0.9%	-3.6%	16	16	16
CA	0.6%	0.9%	4.4%	19	16	11	0.2%	0.5%	-4.7%	19	20	11
CH	1.6%	1.3%	9.4%	14	13	5	0.9%	2.7%	-8.8%	14	14	7
DE	2.4%	2.0%	3.6%	9	9	13	2.0%	8.0%	10.0%	7	7	6
GB	0.9%	0.8%	0.0%	16	17	20	1.1%	3.0%	4.9%	11	12	9
IT	0.6%	0.3%	7.6%	18	19	7	0.2%	0.7%	-1.4%	18	18	20
NL	0.9%	1.0%	0.3%	15	15	17	0.7%	1.9%	-10.3%	15	15	5
NO	0.8%	0.4%	7.5%	17	18	8	0.3%	0.6%	-3.7%	17	19	14
RU	2.2%	1.1%	13.4%	10	14	3	0.2%	0.7%	-1.7%	20	17	19
US	2.1%	2.3%	7.8%	11	7	6	1.0%	3.4%	4.1%	13	11	12
spring	1.9%	1.3%	0.1%	12	12	18	1.1%	3.5%	-4.8%	12	10	10
summer	1.9%	1.7%	1.6%	13	11	16	1.3%	2.8%	-3.1%	10	13	17
autumn	2.5%	1.9%	3.9%	7	10	12	1.3%	4.0%	5.0%	9	9	8
winter	2.5%	2.2%	5.2%	8	8	10	1.6%	4.6%	2.8%	8	8	18
age	24.9%	29.8%	-11.6%	1	1	4	30.6%	93.3%	12.0%	1	1	4
Male	3.8%	4.4%	6.4%	6	6	9	3.3%	15.2%	-4.1%	6	5	13
mood	9.1%	11.4%	-18.4%	4	4	1	7.4%	24.6%	-24.0%	4	4	2
arousal	6.7%	8.3%	0.1%	5	5	19	4.7%	12.5%	12.4%	5	6	3
stress	17.4%	16.5%	17.8%	2	3	2	27.3%	48.8%	38.4%	2	2	1
concentration	16.7%	17.3%	-2.3%	3	2	15	14.4%	29.8%	-3.7%	3	3	15

Table 3.6: Feature importances of the Gradient Boosting Machines (both classifier and regressor) of univariate features with the two targets *momentary tinnitus* and *tinnitus loudness*. To get a better estimate of the feature importance, three different methods have been used: Gini importance, permutation importance, and correlation. The Gini importances within one column add up to 100 %, the permutation importance indicates the absolute increase of the error rate if that feature was left out. Since the percentages cannot be compared between columns, but only within a column, the ranks of the feature importances are also given. The greener a cell is, the more important the feature for the target (*momentary tinnitus* or *tinnitus loudness*) is. The features themselves are grouped in countries, seasons, demographics, and daily questions. As *age* is a feature with high cardinality, it clearly helps the tree-based Gradient Boosting Machines to predict the targets. The high feature importance for the variable *age* could also be an indication of an overfitting of users who have completed very many assessments.

If we firstly divide the features into their groups (country, season, demographics, daily EMA-questions), we can see that the EMA features (questions 4 (mood), 5 (arousal), 6 (stress), and 7 (concentration)) and the demographic features (sex, age) seem to be the most important feature groups on average. The third most important feature group is the season, followed by the countries. *Age* is a very important feature for the Gradient Boosting Machine for two reasons. First, it has a high cardinality (many different values) and second, it has a moderate correlation with current tinnitus. The permutation importance of 29.7 % suggests that the accuracy becomes 29.7 % percentage points worse when the *age* is replaced by a random variable. For example, almost all Russian users have consistently answered the question about current tinnitus in the affirmative. Within the countries feature, Russia therefore has a high correlation with current tinnitus. However,

because there are relatively few users compared to all users, the Gini importance for RU only shows a value of 2.19 %.

3.2.3.2 | RQ2: Is the reported loudness of TYT users predictable using the same features like in RQ1?

In the second research question, we want to estimate the *tinnitus loudness* based on the features, which are listed in Table 3.5. This machine learning task is not a classification, but a regression. Therefore, we tried to optimize the Gradient Boosting Regressor for absolute deviation from the estimated loudness to the true loudness. We refer to this measure as *abs_mean_error*. In contrast to *momentary tinnitus*, there was no skewed distribution with respect to *tinnitus loudness*. This did not, in our estimation, produce a need to generate samples to produce, for example, a Gaussian distribution of the true values. We chose to train the regressor on the mean absolute percent error rate because this measure directly gives a sense of how well or poorly the regressor is performing. In each case, the regressor underestimates the marginal regions (< 0.2 and > 0.7) of the reported loudness and slightly overestimates the middle regions. On average, it is off by 8 percentage points. Thus, if a user reports a loudness of 70 %, the regressor estimates a loudness of 62 - 78 % on average. A density distribution of the reported loudness and the estimated loudness is given in Figure 3.9.

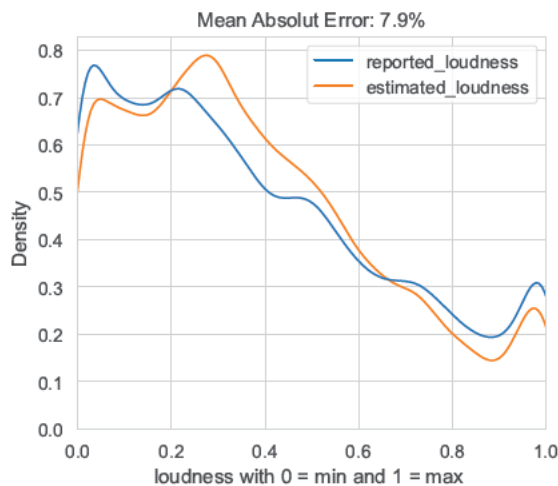


Figure 3.9: Density curves for the reported loudness and the estimated loudness for all assessments of the test set. Users in the marginal areas tend to be underestimated by the regressor (loudness from 0.0 to 0.2 and 0.7 to 1.0). In the middle ranges (loudness from 0.2 to 0.7), they tend to be overestimated. Nevertheless, an overall performance with a mean absolute error of 7.9%-points was obtained.

3.2.3.3 | RQ3: Are we able to reveal country- and season-specific differences for the reported *momentary tinnitus* based on the daily questionnaire of TYT users?

To answer this question, there are 97,742 responses from 3,691 users from a total of 111 countries for the period from April 2014 to February 2021. For the further analysis, we restricted ourselves to the countries at least represented by more than 30 users with more than 300 questionnaires in total. For this subset, with 15 countries, 3,163 users remain with a total of 88,049 filled out daily questionnaires. Most responses are from Germany, with 51,804 completed questionnaires, generated by 1,410 users, whereas the fewest completed questionnaires come from the Federative Republic of Brazil, with 334 completed questionnaires, generated by 50 users. The mean number of filled out questionnaires per country is 5870 (std = 13,058). The mean number of users is 210 (std = 357). For the question of interest, *Did you perceive the tinnitus right now?* (*question1*), mean for 'Yes' is 78.97 % (std = 12.21 %), an interquartile range of 15.73 %, with a maximum value of 95.58 % from Italy, and a minimum value of 48.66 % from Norway, were found.

RQ3₁: Are there country-specific differences for the *momentary tinnitus*? A chi-square test of independence showed that there are significant differences between the countries, $\chi^2(14, N = 85933) = 2441.44, p < .001$. 105 post-hoc χ^2 tests were performed to compare pairwise differences. Using corrected p-values, 91 pairs of countries were rejected ($p = .05$). 14 pairs could not be rejected at $p = .05$, i.e., the pair Germany-Great Britain, and Germany-Sweden. This indicates that these countries have a similar pattern in *momentary tinnitus* occurrence. A detailed overview of the answers of *question1* (*Did you perceive the tinnitus right now?*) is given in Table 3.7. To ensure comparability between the countries under consideration, we have looked at the demographic variables in detail in Table 3.8.

RQ3₂: Are there season-specific differences for the *momentary tinnitus*? To answer this question, we again analyzed only countries represented by more than 30 users with more than 300 completed questionnaires *per season*. This filter setting holds True for Switzerland, Germany, the United States, Great Britain, and the Netherlands. The largest sample is again for Germany, with 51,534 completed questionnaires, the smallest sample is for the UK, with 3,684 completed questionnaires.

If we do not group by country, it can be seen that the greatest probability for *momentary tinnitus* is in summer with 83.4% (std = 8.6%). In contrast, the lowest probability for *momentary tinnitus* is in winter, with 71.0 % (11.8 %). The interquartile range is 14.5 % for winter, and 11.8 % for summer. If we group by country, the highest probability

Country_Name	No	Yes	n_questionnaires	n_users
Australia	14.5%	85.5%	666	77
Austria	29.6%	70.4%	1321	68
Belgium	28.6%	71.4%	972	44
Brazil	8.7%	91.3%	344	50
Canada	13.9%	86.1%	2341	126
France	16.6%	83.4%	467	72
Germany	21.0%	79.0%	51804	1410
Italy	4.4%	95.6%	1220	81
Netherlands	33.1%	66.9%	7268	180
Norway	51.3%	48.7%	1178	42
Spain	9.3%	90.7%	517	82
Sweden	18.2%	81.8%	362	38
Switzerland	32.8%	67.2%	5139	122
United Kingdom	20.5%	79.5%	3713	210
United States of America	12.8%	87.2%	10737	561

Table 3.7: *Momentary tinnitus* by country for individuals of the TYT platform grouped by country. When filling out a questionnaire, most users state that they perceive the tinnitus at that moment. The chance for this is 78 %, with a standard deviation of 12 percent.

for *momentary tinnitus* is in summer in Great Britain (95.7 %), the lowest in winter in Switzerland (60.7 %). The ratios of yes-no-responses are shown in Fig. 3.10. Considering not only these five countries, but all 111 countries in the present data set without setting a questionnaire or user threshold, the probability of *momentary tinnitus* perception is 80.6 % in summer, 80.1 % in fall, 78.6 % in spring, and 75.1 % in winter. A χ^2 test of independence showed that there was a significant association between *season* and *momentary tinnitus* for all countries without a user or questionnaire threshold, $\chi^2(3, N = 95446) = 216.19, p < .001$. Overall user reporting for tinnitus is thus most likely in summer.

Baseline characteristics from this questionnaire for the five countries (CH, DE, GB, NL, US), as well as all other countries, can be seen in Table 3.8. These five countries are the subject of our RQ3₂. To ensure comparability between countries, we considered other demographic data in more detail. For the characteristics *handedness* and *family history of tinnitus complaints*, a χ^2 test was performed. The χ^2 test showed that there was no significant association within the country groups, $\chi^2(8, N=2319) = 6.64, p=0.58$ for *handedness*, and $\chi^2(4, N=2314) = 4.33, p=0.36$, for *family history*. To compare the age distributions between the countries, a one-way ANOVA was performed with $F(4, 2267) = 5.17, p < 0.001$. A post-hoc pairwise Tukey test revealed differences between DE and US (mean diff. = 2.36, $p < 0.05$), and GB and US (mean diff. = 5.07, $p < 0.01$). The remaining eight pairwise groups had no significant differences in their means.

In a slightly different approach, we considered months instead of seasons. Therefore, we increased the granularity of the x-axis. In addition, we examined the respective average temperature per month in relation to tinnitus occurrence for the countries considered

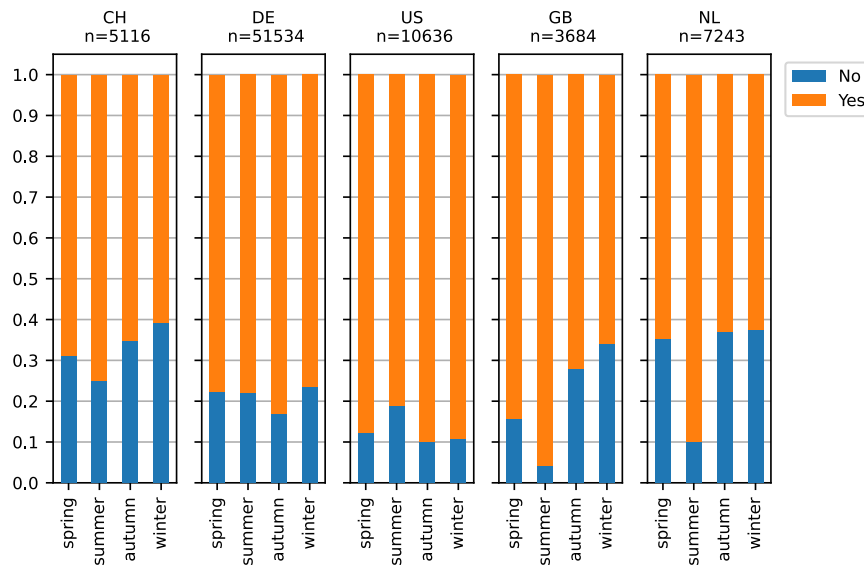


Figure 3.10: Distribution of the *momentary tinnitus* (*Did you perceive the tinnitus right now?*) by country and season for Switzerland (CH), Germany (DE), the United States of America (US), the United Kingdom of Great Britain & Northern Ireland (GB), and the Netherlands. *n* denotes the number of filled out daily questionnaires per country for all seasons.

Country	Sex	Count	Age F(4, 2267) = 5.17, p < 0.001							Handedness X ² (8, N=2319) = 6.64, p=0.58			Family History X ² (4, N=2314) = 4.33, p=0.36	
			Mean	Std	Min	25%	50%	75%	Max	Left	Both Sides	Right	No	Yes
CH	Female	32	48.38	13.84	31	37	47	62	74	0.0%	9.1%	90.9%	69.7%	30.3%
	Male	78	49.94	13.95	21	39	50	59	78	12.5%	17.5%	70.0%	71.3%	28.8%
DE	Female	414	44.36	13.80	8	33	46	55	79	10.5%	13.1%	76.4%	74.3%	25.7%
	Male	851	49.15	13.83	10	39	50	58	87	10.6%	13.1%	76.3%	79.3%	20.7%
GB	Female	91	41.81	12.33	17	32	42	51	70	8.8%	15.4%	75.8%	74.7%	25.3%
	Male	106	46.12	13.13	13	37	46	57	71	13.2%	7.5%	79.2%	78.5%	21.5%
NL	Female	25	50.76	12.07	29	43	47	61	73	5.9%	14.7%	79.4%	73.5%	26.5%
	Male	95	45.79	14.12	18	34	50	57	73	14.0%	8.1%	77.9%	73.5%	26.5%
US	Female	242	47.71	13.19	12	38	49	57	84	14.9%	8.9%	76.2%	69.6%	30.4%
	Male	284	51.58	12.68	16	43	54	60	81	11.5%	12.8%	75.7%	78.9%	21.1%
all*	Female	1102	44.46	13.60	8	33	45	55	84	11.2%	13.4%	75.3%	72.7%	27.3%
	Male	2231	47.15	13.95	1	37	48	57	114	12.9%	15.7%	71.4%	78.2%	21.8%

Table 3.8: Statistical comparison of the five countries CH, DE, GB, NL, and US with all users. Additionally, the data is grouped by gender. For the χ^2 tests, the N differs from the Count column, as some data is missing. The χ^2 for *handedness* and *family history* is not significant. For the comparison of the age distributions, the post-hoc Tukey test shows significant mean differences for Germany with the United States ($p < 0.05$), and Great Britain with the United States ($p < 0.01$). The table supports the comparability of the five countries that are mainly discussed in RQ3. *The five countries CH, DE, GB, NL, and US are included in *all* countries.

(i.e., Switzerland, Germany, U.S., Great Britain, and the Netherlands). When multiple temperature data points from different cities were available for a country, they were

aggregated with the average.

In this context, a positive correlation means the higher the temperature, the more likely is the *momentary tinnitus*. A high positive correlation can be obtained for the Netherlands ($r(10) = .83, p < .001$), for Great Britain ($r(10) = .86, p < .001$), and for Switzerland ($r(10) = .72, p = .009$). On the contrary, the U.S. shows a non-significant medium negative correlation ($r(10) = -.41, p = .18$). For Germany, however, the correlation between temperature and tinnitus occurrence can be considered uncorrelated ($r(10) = -.09, p = .78$). The cyclical temperature pattern associated with tinnitus over the year for the various countries is shown in Fig. 3.11. There was a statistically significant difference between the countries as determined by one-way ANOVA ($F(4, 55) = 6.69, p < .001$). A post-hoc Tukey test indicates that the annual course of *momentary tinnitus* is different between the country pairs Netherlands-U.S. ($p < .01$) and Switzerland-U.S. ($p < .01$).

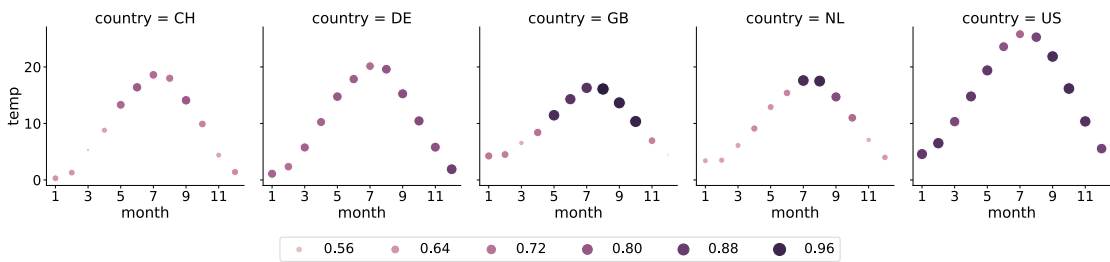


Figure 3.11: Cyclical temperature pattern associated with tinnitus for Switzerland (CH), Germany (DE), the United States of America (US), the United Kingdom of Great Britain & Northern Ireland (GB), and the Netherlands. The x-axis shows the month, the y-axis the temperature in degrees Celsius. The larger the circle is, the higher the average probability for a *momentary tinnitus* for this country in this month is. The size and color of the cycles indicate the chance of *momentary tinnitus*. The bigger the cycle, the higher the chance.

RQ3₃: In the light of a combination of country- and season-specific differences, the question arose, whether the *momentary tinnitus* varies within the year and across countries. In contrast to the previous section, we have ignored temperature in this question. Instead, we examined the following: For each of the countries considered, and for each individual month of the year, we calculated the probability of tinnitus by dividing the number of yes responses by the sum of responses. In the following step, we examined the probability of tinnitus over the course of the year. To increase comparability, we additionally calculated the average of the tinnitus probability for all available data on a monthly basis.

Since most of the data comes from Germany, this country has a correspondingly large influence on the average values. Accordingly, the curve for Germany is very similar to

the curve of all data (statistic = .17, $p = 1.00$). On the contrary, the Netherlands, the U.S., and Switzerland reveal a different distribution of the tinnitus with p -values < 0.01 . For Great Britain, the distribution can be considered to be slightly different as p -value is .10. An overview of the distributions compared with the average is given in Fig. 3.12. A summarizing statistical overview, in turn, is given in Table 3.9.

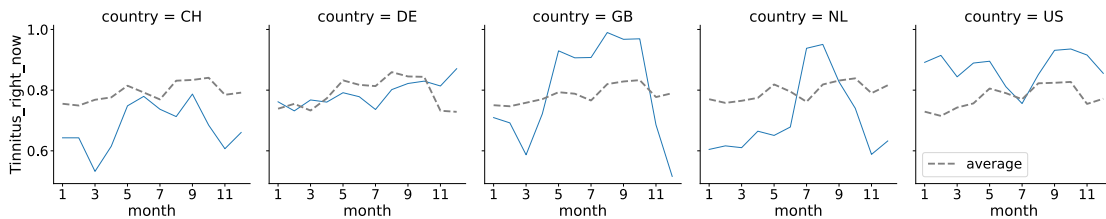


Figure 3.12: Course of occurrence of tinnitus over the year for Switzerland (CH), Germany (DE), the United States of America (US), the United Kingdom of Great Britain & Northern Ireland (GB), and the Netherlands. The x-axis shows the month, the y-axis the probability for tinnitus occurrence. The dashed grey lines show the average of tinnitus occurrence for all data *except* the country plotted on this axis. The graph indicates that people of different nations perceive tinnitus differently throughout the year.

country	count	mean	std	min	25%	50%	75%	max
CH	12.00	0.68	0.08	0.53	0.64	0.67	0.74	0.79
DE	12.00	0.79	0.04	0.73	0.76	0.78	0.82	0.87
GB	12.00	0.80	0.16	0.52	0.69	0.81	0.94	0.99
NL	12.00	0.71	0.13	0.59	0.61	0.66	0.76	0.95
US	12.00	0.87	0.05	0.76	0.85	0.89	0.91	0.94

Table 3.9: Statistics for the occurrence of tinnitus throughout the year grouped by country. *Count* simply represents the number of months in a year. For this data set, *momentary tinnitus* occurred least in Switzerland in March (53 %), and most in the UK in August (98 %).

The highest probability for tinnitus is in the US with an average chance of 87 %, the lowest probability in Switzerland with 68 %. The largest variance occurs in Great Britain, with 16 % standard deviation, the smallest in Germany, with 4 %. For this data set, tinnitus occurred least in Switzerland in March (53 %), and most in the UK in August (98 %).

RQ3₄: The question arose, whether country- and season-specific differences of the reported worst symptom can be identified. To answer this research question, we again focused on the five countries [CH, DE, GB, NL, US]. When registering on the TYT

platform, the question about the worst tinnitus symptom is asked once. For each country and season, we calculated the relative number of answers within a country to compare which symptom is more likely in which season. Each column adds up to 100 %. The 1,310 users from Germany had the lowest standard deviation (.94 std). The Netherlands with 175 users had the largest standard deviation (2.01 std). *I find it harder to relax* is the most likely symptom in the Netherlands in fall, with 8.57 %, and, at the same time, with a global maximum. *Feeling depressed* ranks second for the UK and the Netherlands. For the U.S., the two worst symptoms are *difficulty following a movie or conversation* and *concentration problems*. For the U.S., however, there is little variation between seasons within these two worst symptoms. *None of these symptoms* ranks second for Switzerland. *Irritability with friends and family* is the least indicated worst symptom for all countries. A chi-square test was performed between distribution of the worst symptom and country. There was no statistically significant relationship between worst symptom and country, $\chi^2(40, N=6) = 0.53, p=1.0$.

In a similar approach, we disregarded countries and investigated the evolution of the worst tinnitus symptom between seasons. Thus, we examined whether there are different worst symptoms per season. *Because of the tinnitus I am more irritable with my family, friends and colleagues* is the most unlikely symptom (mean = 5.9 %, std = 1.0 %). The most likely symptom constitutes *I find it harder to relax because of the tinnitus* (mean = 17.7 %, std = 1.9 %). Details are given in Fig. 3.13. Difficulties in relaxing is the worst symptom across all seasons. The data further indicates that feelings of depression are stronger in the months of autumn and winter. Difficulties in following conversations are more pronounced in summer. Irritability with colleagues or family is the least selected symptom. However, a chi-square test of independence showed that there was no significant association between worst symptom and season, $\chi^2(24, N = 3458) = 30.86, p = .16$.

3.2.4 | Discussion

The present work investigated the differences of *momentary tinnitus* and *tinnitus loudness* in relation to seasons and countries.

- To summarize the results for RQ1, we found that we can predict the *momentary tinnitus* with an accuracy of 94.03 % on the assessment level.
- For RQ2, we found that the *tinnitus loudness* can be regressed with a mean absolute error rate of 7.9 %-points on a scale from 0 to 100 %.
- For RQ3₁ (country specific differences for the *momentary tinnitus*), we found that most of the countries report the *momentary tinnitus* differently.

worst_symptom	season	CH (n=114)	DE (n=1310)	GB (n=201)	NL (n=175)	US (n=537)
Because of the tinnitus I am more irritable with my family, friends and colleagues.	spring	0.00%	1.91%	0.50%	1.14%	0.93%
	summer	0.00%	1.37%	1.00%	1.14%	1.86%
	autumn	1.75%	1.68%	0.50%	2.29%	2.23%
	winter	0.88%	1.53%	1.00%	0.57%	0.93%
Because of the tinnitus I am more sensitive to environmental noises.	spring	4.39%	2.67%	1.99%	1.14%	1.49%
	summer	0.88%	1.83%	1.00%	1.71%	2.23%
	autumn	4.39%	2.90%	0.50%	4.00%	2.61%
	winter	2.63%	2.14%	1.00%	0.00%	1.68%
Because of the tinnitus it is difficult to concentrate.	spring	0.88%	3.44%	2.49%	2.86%	4.66%
	summer	2.63%	2.90%	1.99%	1.14%	2.79%
	autumn	0.88%	3.66%	1.49%	6.29%	5.21%
	winter	1.75%	2.60%	1.00%	3.43%	2.79%
Because of the tinnitus it is difficult to follow a conversation, a piece of music or a film.	spring	2.63%	3.36%	3.48%	1.14%	4.28%
	summer	2.63%	2.98%	5.97%	3.43%	4.66%
	autumn	3.51%	4.12%	2.49%	3.43%	3.17%
	winter	2.63%	3.59%	2.49%	1.71%	3.91%
Because of the tinnitus it is hard for me to get to sleep.	spring	4.39%	2.60%	2.99%	1.14%	3.54%
	summer	1.75%	1.98%	2.99%	0.57%	2.98%
	autumn	0.88%	3.36%	4.98%	5.14%	3.17%
	winter	3.51%	2.67%	5.47%	1.71%	4.10%
I am feeling depressed because of the tinnitus.	spring	3.51%	1.91%	2.99%	3.43%	0.93%
	summer	0.88%	2.14%	4.48%	2.86%	2.05%
	autumn	4.39%	2.14%	4.48%	5.14%	3.91%
	winter	1.75%	1.60%	6.47%	2.86%	2.79%
I don't have any of these symptoms.	spring	6.14%	2.67%	1.00%	0.00%	1.68%
	summer	1.75%	2.37%	1.00%	1.71%	1.68%
	autumn	4.39%	3.05%	1.00%	2.29%	2.42%
	winter	7.89%	2.37%	1.99%	2.29%	2.42%
I find it harder to relax because of the tinnitus.	spring	4.39%	5.19%	6.97%	5.71%	4.47%
	summer	7.89%	3.44%	2.49%	4.57%	3.54%
	autumn	3.51%	5.57%	4.48%	8.57%	3.91%
	winter	3.51%	3.28%	7.96%	2.86%	2.23%
I have strong worries because of the tinnitus.	spring	3.51%	2.21%	2.99%	0.00%	2.79%
	summer	0.88%	2.60%	1.49%	4.57%	1.49%
	autumn	1.75%	3.59%	3.48%	6.29%	1.68%
	winter	0.88%	2.60%	1.49%	2.86%	2.79%

Table 3.10: Distribution of the worst symptom for each country and season. We only considered countries with more than 300 questionnaires from more than 30 users. Each column adds up to 100 %. n denotes the number of users from this country.

- Furthermore, for RQ3₂, we also found season-specific differences for the momentary tinnitus. If we do not group the data by country, *momentary tinnitus* is most likely to occur in the summer. This is in contrast to the results of [208]. When we group our data by country, an ambiguous picture emerges between countries for the most likely season for tinnitus.
- Regarding RQ3₃, we found that the *momentary tinnitus* does vary within the year and within one country. We also found that this *momentary tinnitus* variance within one country is different to other countries, i.e., if we compare Great Britain with the US.

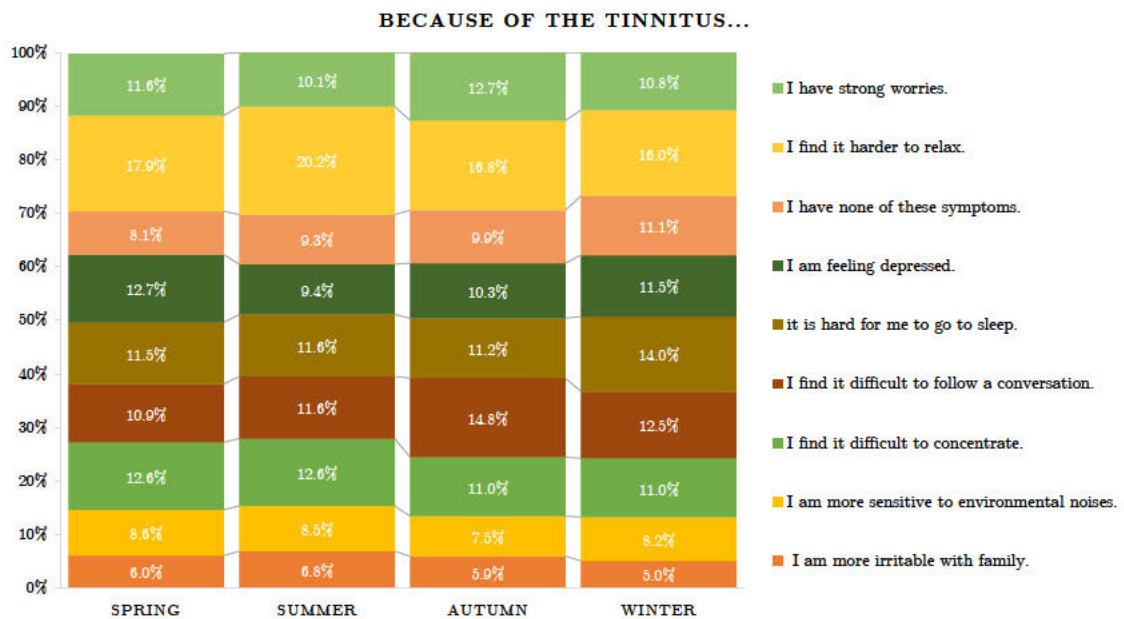


Figure 3.13: Development of the worst symptom for tinnitus over the seasons. For countries in the southern hemisphere, the seasons have been inverted. Users are asked this question once when completing the baseline questionnaire (n = 3458). *Irritability with friends and family* is the least selected symptom, *difficulty with relaxation* is the most selected. Difficulty following a conversation has a clear high in summer. The values for each season add up to 100 %. A chi-square test of independence showed that there was no significant association between worst symptom and season, $\chi^2(24, N=3458) = 30.86$, $p=.16$.

- For RQ3₄, we examined whether the distribution of the worst symptom changes between years or whether it is significantly different between countries. Our analysis showed that neither varied significantly, although the numbers suggest small differences.

Although we found significant differences for the *momentary tinnitus* between seasons and countries, this does not establish causality between the features and the target. Additionally, although our findings potentially provide important insights for further tinnitus research, there are a few limitations that should be discussed. First, there might be a myriad of other reasons why tinnitus is more likely in some countries in summer and in some in winter. Influencing factors could be, for example, air pressure, stress level, or the number of hours of sunshine. Second, user numbers vary widely between countries. This can lead to a selection bias in the evaluation. Consider the filter criterion "at least 30 users per country". If one user was particularly active in filling out the daily questionnaire, and the other 29+x users were not, this

might lead to a selection bias. Third, although our research results indicate different seasonal trends for the *momentary tinnitus* for different countries, there may be individuals who perceive tinnitus seasonally quite differently, possibly even completely in the opposite direction. This means that these findings are not applicable to individuals.

For the worst tinnitus symptom per country and season, comparability between countries and seasons may also be biased by the selection due to the low number of users per category. For Switzerland, for example, we would expect 3.17 individuals per symptom per season (i.e., 2.8 % per line), if symptoms and seasons were equally distributed. In this respect, it is surprising for Switzerland, for example, that *relaxation* is more difficult in summer (7.89 %) than in winter (3.51 %). The situation is different with Germany. Here, we have a large number of users of 1,310 and would expect 36.4 individuals per category, if the symptoms were equally distributed among all seasons. This argument is supported by the fact that the variance in Germany is lower than in Switzerland. Nevertheless, we can observe for Germany that *relaxation* is more difficult for spring and autumn (about 5 %) than for summer or winter (about 3 %).

The accuracy depends on which level we split: Assessment level vs user level. By stratifying at the **assessment level** (i.e., on the level of filled out questionnaires), one can ensure that the distribution of the target between test and training data remains the same. The specific problem at hand is that several users have filled out different numbers of assessments. There are many users with only one or two assessments, and a few users with several hundred or thousand assessments (so-called power users), as you can see in Fig. 3.14. These power users are highly likely to be present in the training, validation, and testing data. Any model is therefore predestined to an overfitting on these power users. One can address this problem by excluding users that are in the training set from the test set and vice versa. We then no longer evaluate at the assessment level, but at the **user level**. However, the accuracy in the test set then drops from 94 % to around 50-60 %, depending on different test sets with different power users. That is, the model can hardly predict assessments from users it has never seen before and should be considered for practical implications.

There are features that are user-dependent and therefore reduce the number of learnable parameters in the model when splitting the data at user level. These include, for example, country, gender, age, and season. If by chance there are only German users in the training data, but English users in the test data, then the feature *country* has no more variance and therefore no prediction power for the model. As another example, if a male user who is 43 years old reports the *momentary tinnitus* as "Yes", several hundred times, then

the model learns that 43 year old males always have tinnitus. However, this would have nothing to do with the dynamic assessments and therefore contradicts the idea of Ecological Momentary Assessments. This would partially explain the drop in accuracy between the training and test sets. We therefore took a subset of the features that we know retain their variance, even when split at the user level. These features are *mood*, *arousal*, *stress* and *concentration*. If we now split at the assessment level, i.e., allow the same users in the training and test data, we get an accuracy of 84 % in the test set, which is significantly better than guessing. If we now additionally split on user level, the accuracy drops again to 50-60 %, which suggests an overfitting of the training users. Thus, the model cannot predict assessments of users that it has not yet seen, or to put it in another way: The completion behavior of the individual user varies so much between the users that one can hardly conclude from user A to user B.

The bias in the selection of users remains: A user who has completed many assessments is represented in both the train and test data, which raises doubts about the generalizability of the model, since one may have trained a user-specific model. On the other hand, if one tries to stratify for users, the distribution of the target, and demographic data, no more data remain and one would have to collect a large amount of more data, which is expensive and time consuming. Any stratification technique eventually creates a bias. We decided for the user bias to be able to stratify correctly for the target. This also allowed us to use more data to train our models. The generalizability of the model to users from a different population is not known. However, it is known that the model can make predictions at the assessment level for users who come from a known population. This is shown by the high accuracy of the test set at the assessment level. In current investigations, we evaluate these differences more in-depth.

Feature Importance High cardinality features such as *age* and the daily questions are assigned with a higher importance as these features can be easily split up into multiple, potentially pure subsets. For binary features, the tree classifier can only split up the data once. However, for features with high cardinality, the tree can potentially split up the data $n_{unique} - 1$ times. Feature importance does not establish causality between input variables and target. It is rather an estimator of which variable has the greatest predictive power for the Gradient Boosting Machine. Any other classifier, such as a neural network, would potentially produce a different ranking for feature importance. Among the percentages, the 93.3 % permutation importance for *age* in the regressor model is prominent. The 93.3 % induces that the model loses almost all its predictive power without the *age* feature. However, since the model was trained and evaluated with

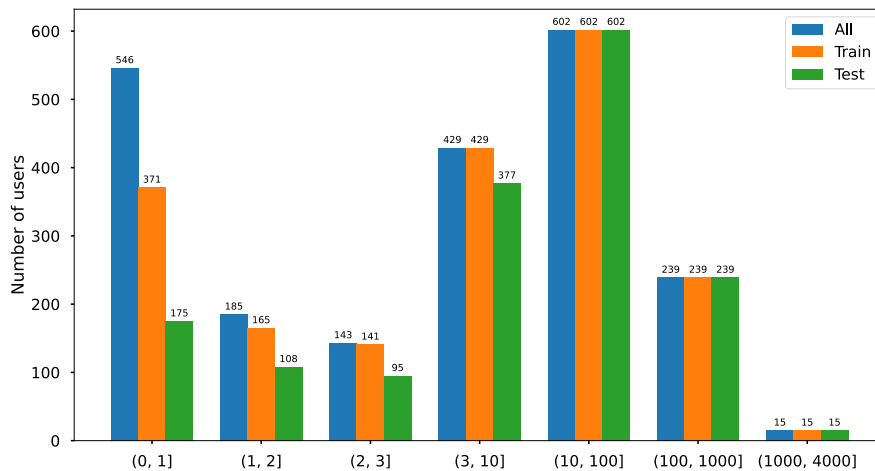


Figure 3.14: Number of users by range of filled out questionnaires. If we take only users that filled out one single assessment, we also automatically split on a user level. That means, the model predicts only on users it has never seen before. However, if we include users that have filled out the questionnaire more than 10 times, the likelihood that users are represented in both the train and test set is very high.

the *mean absolute error*, this percentage value cannot be easily transferred to the mean absolute error but is only an indicator for the importance for the model.

The temperature dataset Although the more than 300 different sources of the individual figures are very well referenced within Wikipedia, there could be noise in the data because, for one thing, only a few cities in a country are a limited representation of the temperature across the country. Second, noise may occur because the temperatures come from different years, some of which were also recorded before the EMA data were collected. Nevertheless, we believe in the reliability of the temperature data because temperatures hardly change significantly within a decade in a country.

Worst Season for Tinnitus We define one season as worse than another if the probability of the *momentary tinnitus* is higher on average. This question cannot be answered unambiguously and conclusively. Related work on tinnitus and seasonality does suggest winter as the worst season [208; 210; 211]. However, 41.8 % of individuals (n = 100) report perceiving summer as the second worst season, which argues against the theory of seasonal affective disorders [75]. In the study, which aggregated tinnitus search requests from online platforms by season and country, the winter was also highlighted as a more frequent season. In the study, which aggregated tinnitus search requests from online platforms by season and country, the winter was also highlighted as a more frequent season

[208]. However, the results are different, even for countries with similar longitudes. For example, this is the case for Sweden and the United Kingdom. The noise in the results could be due to confounders, or the mentioned selection bias.

Outlook In future work, we are heading into two research directions. At first, we plan to compare the results of TYT to other data sources that have similar characteristics. Second, a more in-depth inspection of the user- and assessment perspective of TYT in particular will be considered.

3.2.5 | Data availability

According to the GDPR, the data to replicate these results are available upon request to the corresponding author. Any code to replicate the results, numbers, figures, and tables is publicly available on github.com/joa24jm/tinnitus-country.

Supplementary Information

The Python code to replicate the Machine Learning classifiers, figures and tables is available on github.com/joa24jm/tinnitus-country.

3.3 | Self-Assessment of Having COVID-19 with the Corona Check mHealth App

- **Authors** | Beierle, Felix; Allgaier, Johannes; Stupp, Carolin; Keil, Thomas; Schlee, Winfried; Schobel, Johannes; Vogel, Carsten; Haug, Fabian; Haug, Julian; Holfelder, Marc; Langguth, Berthold; Langguth, Jana; Riens, Burgi; King, Ryan; Mulansky, Lena; Schickler, Marc; Stach, Michael; Heuschmann, Peter; Wildner, Manfred; Greger, Helmut; Reichert, Manfred; Kestler, Hans; Pryss, Rüdiger
- **Published in** | IEEE Journal of Biomedical and Health Informatics, 27(6), 2794-2805, 2023.
- **Available at** | <https://pubmed.ncbi.nlm.nih.gov/37023154/>

Abstract

At the beginning of the COVID-19 pandemic, with a lack of knowledge about the novel virus and a lack of widely available tests, getting first feedback about being infected was not easy. To support all citizens in this respect, we developed the mobile health app Corona Check. Based on a self-reported questionnaire about symptoms and contact history, users get first feedback about a possible corona infection and advice on what to do. We developed Corona Check based on our existing software framework and released the app on Google Play and the Apple App Store on April 4, 2020. Until October 30, 2021, we collected 51,323 assessments from 35,118 users with explicit agreement of the users that their anonymized data may be used for research purposes. For 70.6% of the assessments, the users additionally shared their coarse geolocation with us. To the best of our knowledge, we are the first to report about such a large-scale study in this context of COVID-19 mHealth systems. Although users from some countries reported more symptoms on average than users from other countries, we did not find any statistically significant differences between symptom distributions (regarding country, age, and sex). Overall, the Corona Check app provided easily accessible information on corona symptoms and showed the potential to help overburdened corona telephone hotlines, especially during the beginning of the pandemic. Corona Check thus was able to support fighting the spread of the novel coronavirus. mHealth apps further prove to be valuable tools for longitudinal health data collection.

3.3.1 | Introduction

At the beginning of the COVID-19 pandemic, there was a lot of uncertainty about the novel coronavirus SARS-CoV-2 and knowledge about it was sparse. Quickly, healthcare

systems were overstrained, and telephone hotlines overburdened. There was a huge demand for getting a quick first assessment about the probability of being infected and support in case of a possible infection.

We designed and developed an mHealth (mobile health) app called Corona Check that allows users to answer a questionnaire to get a first assessment about their symptoms/situation with regards to being infected with SARS-CoV-2. We released Corona Check on April 4, 2020, on Google Play and the Apple App Store. As of October 30, 2021, there were almost 90 thousand assessments from more than 50 thousand users. While similar systems have been proposed before [216; 217; 218], to the best of our knowledge, we are the first to report about a large-scale deployment.

The first main benefit of our system was that it gave users immediate individualized feedback and behavioral recommendations based on their symptoms and contact history. Moreover, the users received general hygiene behavior tips. The system thus alleviated pressure from, e.g., telephone hotlines. The second main benefit of Corona Check was that, with the consent of the users, we were able to collect age- and sex-specific data for research about the occurrence and regional spread of corona symptoms.

Our main contribution is the introduction of our system Corona Check, highlighting how a quickly developed mHealth system aimed at the average user can support vast amounts of users in the beginning as well as during the COVID-19 pandemic. In section 3.3.2, we give an overview about related work. In section 3.3.3, we present the technical details of Corona Check, including the user perspective as well as the system architecture and special requirements for apps in the medical field. Section 3.3.4 presents an overview of the data collected so far and age- and sex-specific results. In section 3.3.5, we discuss the limitations of our system. In section 3.3.6, we discuss our results, draw conclusions, and point out future work.

3.3.2 | Related Work

Information technology and methods from computer science have been used from the very beginning of the COVID-19 pandemic in a variety of ways and for a variety of purposes. With vast amounts of data being quickly available, artificial intelligence and especially machine learning is often applied in COVID-19-related research. In their survey paper, Khan et al. distinguish between diagnosis, screening, prediction of COVID-19, and drug research [219]. Research related to COVID-19 that is employing machine learning, for example, is about analyzing x-ray images [220; 221; 222; 223; 224; 225] and cough sounds [226; 227], or about making predictions about the spread of the virus [228; 229; 230]. What the mentioned research has in common is that the users of

these systems were doctors, epidemiologists, researchers etc. – but not laypersons from the general public.

In contact tracing apps, the average user interacts with a system related to COVID-19. After Alice has been in close physical proximity with someone called Bob who later tested positive for SARS-CoV-2, typically measured by Bluetooth signal strength of smartphones nearby, she can receive a notification and get herself tested. Software system architectures, privacy concerns, and public perception of contact tracing apps have been discussed extensively [231; 232; 233; 234].

Another related category of apps is that of symptom tracking. Menni et al. reported about an app that tracks potential symptoms [235]. A fraction of the app users has undergone a COVID-test, revealing that loss of smell and taste was higher in those who tested positive. Klaser et al. [236] used the same app for tracking levels of anxiety and depression in the UK, finding small associations between SARS-CoV-2 infection and anxiety and depressive symptoms. An updated version of the app was used to track symptoms of infected people to compare symptoms between the delta and omicron variants of SARS-CoV-2 [237].

Much less attention was spent on mHealth or expert systems that regular users can use in order to get a first feedback based on their own symptoms. Especially in the early phases of the pandemic, there was a high level of uncertainty in the population about how to behave if there was a suspicion of infection. At this stage, expert recommendations were offered by telephone hotlines from the public health system. These recommendations were primarily based on symptoms, contact history and travel history, as laboratory tests were not yet widely available.

An app-based expert system which provides individualized recommendations based on symptoms, contact and travel history has many advantages: It can provide important individualized information in an efficient way, can be easily updated in the rapid changing situation (e.g., the emergence of new high-risk areas) and scales easily. By providing easy access to an accurate and up-to-date individualized recommendation, an app-based expert system can importantly contribute to inform the population how to best behave in order to protect themselves and others, which is an essential aspect in the early management of the pandemic. Moreover, the app-based expert system has the potential to relieve pressure of overloaded telephone hotlines or overwhelmed experts.

At the very beginning of the pandemic (publication in March and April 2020), two papers were published sketching the idea of mHealth expert systems for the diagnosis of infections with the novel coronavirus [216; 217]. Both papers present PC software prototypes based on the idea of having a rule-based system and checklists of symptoms

that the user fills out. There are no reports of deployment nor detailed evaluations of the developed prototypes.

Hakim et al. also developed an Android expert system for diagnosing COVID-19 based on a rule-based system [238]. The input is a questionnaire about symptoms and travel and contact history. The authors reported about a small user study with 12 participants. Banjar et al. reported about the development of a prototype of a COVID-19 diagnosis and management expert system [239]. The target audience are doctors in Saudi Arabia. The expert system handles the patients' data, their Electronic Health Records (EHRs), and processes current COVID-19 guidelines in order to classify the patients by their medical condition.

Mufid et al. developed an Android app that consists of an expert system for early detection of having COVID-19 and an information module that displays current news about the spread of the virus [240]. The app was tailored specifically for Indonesia. The expert system is based on a 16-item questionnaire about symptoms and contact with infected people. The system returns a risk status from "very low", over "medium", to "high risk". The authors reported about a usability study with several participants (exact sample size not specified).

A related field to expert systems is chatbots. Battineni et al. developed a chatbot asking the user about symptoms and refer the user to a doctor if a certain threshold is met with the answers. In their evaluation, the authors compared their approach to other existing chatbots. Erazo et al. developed a web-based chatbot to alleviate the pressure on the health care system [218]. A small-scale user study (exact sample size not specified) showed that the users found the system useful. Almalki et al. cover more about chatbots related to the COVID-19 pandemic in their survey paper [241].

There are some works on using wearables or other small devices for trying to detect SARS-CoV-2 infections. Mukhtar et al. developed a device based on Arduino hardware that measures heartbeat, cough severity, temperature, and blood oxygen level for detecting COVID-19 [242]. In contrast to questionnaire-based solutions that the average user can perform on his/her smartphone, this solution is rather targeted at use in hospitals or to monitor patients at home. Astriani et al. presented a smart mirror measuring heart rate and temperature in order to warn about possible infections [243].

There is some work that proposed designs, frameworks, or methodologies for systems that detect an infection with the novel coronavirus. Skibinska et al. proposed a methodology for early-stage detection of COVID-19 based on data from wearables [244]. Maghded et al. proposed a design for a framework that uses the smartphone's sensors to detect an infection with the coronavirus [245]. They proposed using a variety of sensors from the smartphone, including, e.g., measuring the temperature or taking photos of CT scan

images of the lung. Belkacem et al. proposed a hypothetical end-to-end-pipeline for detecting different respiratory infections [246]. What these approaches have in common is that they all rely on study results and/or machine learning models that contain knowledge about SARS-CoV-2 infections gained during the ongoing pandemic. Similarly, Li et al. developed a mobile system capable of analyzing x-ray images of COVID-19 patients [247]. Imran et al. developed an app that records coughing sounds, analyzes them in the cloud, and returns a preliminary diagnosis of COVID-19 [248].

Overall, COVID-19 related research is being conducted in many directions. The amount of related work specifically in the domain of mobile mHealth systems or expert systems is sparse. What the existing related work in the domain has in common is that a lot of the work was preliminary. To the best of our knowledge, we are the first to report about a large-scale deployment of a mobile system in this context.

3.3.3 | Technical Details

In this section, we present the technical details of Corona Check. Corona Check is based on the TrackYourHealth platform and API [249; 250; 251; 252]. The TrackYourHealth platform proved as a valuable base for several other questionnaire-based health-related apps [253; 254]. The backend is based on PHP and serves to native mobile apps for Android and iOS. For technical details of TrackYourHealth, please refer to the cited papers. In this work, we highlight the parts that are specific to Corona Check. In section 3.3.3.1, we show Corona Check from the user perspective, while in section 3.3.3.2, we detail the core functionality of Corona Check and the feedback system that gives the user immediate feedback after filling out the COVID-19 evaluation questionnaire. Section 3.3.3.3 describes the tips module of Corona Check, which provides the user with additional information about the ongoing pandemic. In section 3.3.3.4, we give an overview of the data we collected, and in section 3.3.3.5, we briefly highlight the specific requirements that were necessary for publishing an app related to the coronavirus.

3.3.3.1 | User Perspective

Corona Check was released on Google Play and the Apple App Store in German and English on April 4, 2020. Figure 3.15 shows the general user journey when using the app. (1) The user can start a questionnaire that is filled out for himself/herself or another person. This is shown in Fig. 3.15 on the top left. (2) After starting the questionnaire, the next screen will ask for additional information, for example, as shown in the figure, information about travels. (3) Next, the questionnaire screen is shown where the user answers COVID-19-specific questions and demographic questions. If the user agrees, a

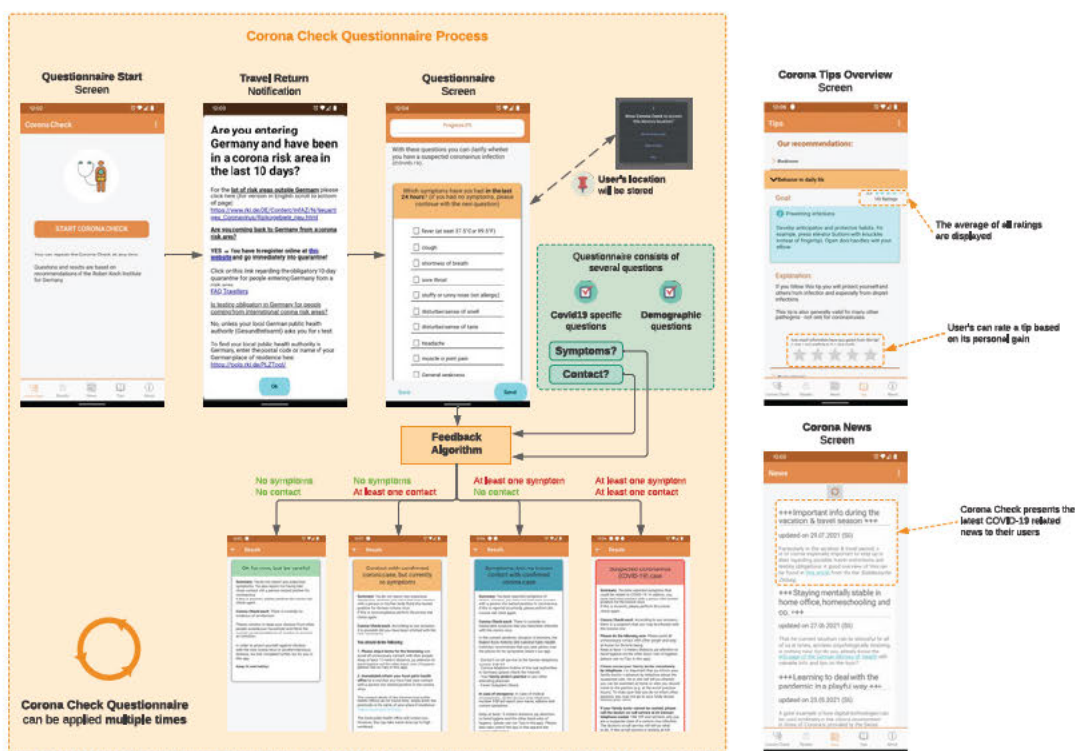


Figure 3.15: Overview of the whole Corona Check process.

coarse location (to protect the privacy of a user) of the device will be stored. Table 3.11 shows the questionnaire used in Corona Check and the answer options. (4) The *Feedback Algorithm* processes the user’s questionnaire input and returns for possible results, see also the bottom of Fig. 3.15. The whole questionnaire process can be repeated at any point in time. Furthermore, users can access the history of past completed questionnaires at any point in time.

Corona Check has two additional features besides the check itself. The *Tips* section (top right of Fig. 3.15) contains health-related tips and recommendations for daily life during the pandemic, e.g., on hygiene. Each tip can be rated with 1 to 5 stars and the average rating is displayed for each tip. The *News* section (bottom right of Fig. 3.15) displays the latest news about the ongoing pandemic.

3.3.3.2 | Feedback System

We developed the questionnaire and rule-based feedback system with the medical and public health experts in our team (including the Bavarian Health and Food Safety Authority). Note that the list of symptoms (Question 1 in Table 3.11) is based on the available

#	Question	Answer Options
1	Which symptoms have you had in the last 24 hours? (if you had no symptoms, please continue with the next question)	"fever (at least 37.5C or 99.5F)", "cough", "shortness of breath", "sore throat", "stuffy or runny nose (not allergic)", "disturbed sense of smell", "disturbed sense of taste", "headache", "muscle or joint pain", "general weakness"
2	Did you have close contact* with a confirmed corona case in the last 14 days before the onset of symptoms OR If none of the above symptoms are present: Have you had close contact* with a confirmed corona case in the last 14 days (as of today)? *Close contact with a confirmed corona virus case is defined as: contact at a distance of less than 2 meters for a total of 15 minutes or more e.g., during a conversation, or if the person lives in the same household or by direct contact with body fluids of this person (e.g., by coughing, sneezing, kissing, contact with vomit, mouth-to-mouth respiration).	"Yes", "No"
3	Age	"0-9 years", "10-19 years", "20-29 years", "30-39 years", "40-49 years", "50-59 years", "60-69 years", "70-79 years", "80 years and older"
4	Sex	"female", "male", "diverse", "no comment"
5	How many years did you go to school (or how many years do you intend to go to school)?	"9 years or less", "10 to 11 years", "12 years or longer", "no comment"
6	For whom are you filling out the questionnaire?	"for myself", "for another person", "no comment"
7	May we use your data for research purposes?	"Yes", "No"

Table 3.11: Questions and answer options in Corona Check.

knowledge at the time. Diarrhea was an additional symptom option to choose during the first two months Corona Check was online. Questions 3, 4, and 5 on age, sex, and education served two purposes for future studies with the Corona Check data. Firstly, these variables can be used to check how representative the Corona Check user group is of a general population. Secondly, we can analyze COVID-19 while controlling for these variables. Educational level was used as a proxy for social status, which has been shown to influence health [255].

The feedback and tips were developed in accordance with the recommendations by the German federal agency for disease control and prevention (RKI, Robert Koch Institute). Overall, there are four possible outcomes bound to the current symptoms and the past contact with infected persons (note that Corona Check was released prior to the availability of vaccinations): (a) no symptoms and no contact, (b) no symptoms and at least one contact, (c) at least one symptom and no contact, (d) at least one symptom and at least one contact (also see Fig. 3.15). The color-coding of the results immediately indicates the level of concern. Each result gives a detailed answer about the situation with advice for the next steps. The feedback algorithm lies at the core of the Corona Check expert system. It fulfills the role that typically humans fulfill at the end of a telephone hotline. Corona Check asked the questions that the human operator would and gives the advice that the person would.

3.3.3.3 | Corona Tips

Corona Check users were provided with 30 different tips¹ on how to safely deal with the pandemic. The 30 tips have been developed by our medical and public health experts. For example, they explain the importance of proper hand washing or complying with contact restrictions. Besides explaining the importance, the tips also contain practical advice for concrete behavior. The tips were displayed unrelated to symptoms entered by the user. All tips were displayed for all users.

3.3.3.4 | Database

The user-generated as well as the operational data of the server application are stored and managed in a relational database, since data integrity and consistency mechanisms (e.g., constraints and ACID transactions) are integrated and well-tested. Moreover, it is easier to create and maintain highly interrelated data models with this type of database system. In addition, relational databases provide a sophisticated query language suitable for

¹Note that in the results section, section 3.3.4, all tips are listed in the context of their evaluation.

complex analytical queries required for data evaluation. To give insights into the project's database schema, Fig. 3.16 depicts an Entity-Relationship Model (ERM) of the Corona Check database using crow's foot notation. Note that Fig. 3.16 illustrates an excerpt of the entire database schema. The selection represents the core entities containing most of the user-generated data. Due to the multilingual project setup, the database schema contains several entities with translated text data, marked with a $\text{\textcircled{T}}$, which are omitted in the ERM to further simplify the model.

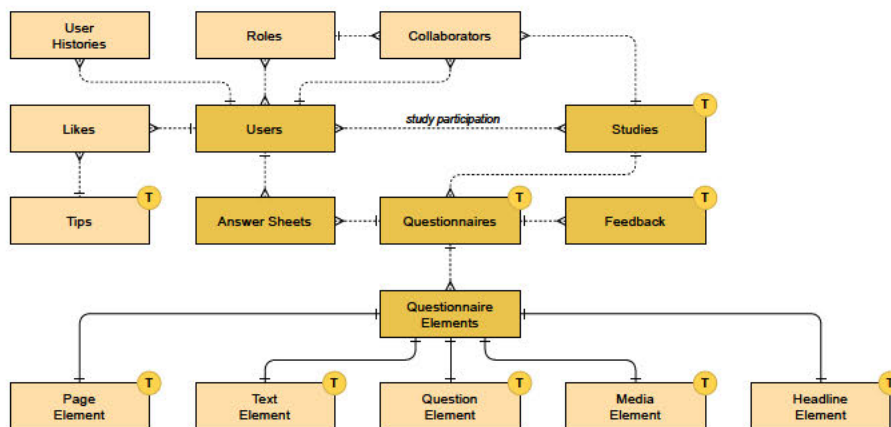


Figure 3.16: Excerpt of Corona Check's relational database schema. The $\text{\textcircled{T}}$ indicates the presence of translated text data.

To describe the data set in this work, a snapshot was extracted on October 30, 2021. Corona Check combines the ideas of mobile crowdsensing [256; 257; 258] and Ecological Momentary Assessments [259] to collect user data. For this reason, the entity `Users` constitutes a central and high-related table with 145,223 verified users. In addition, the users' actions (e.g., questionnaire filled out) are stored in the table `User Histories` to provide insights into the usage of Corona Check. Technically, Corona Check could contain multiple different studies (i.e., (sets of) questionnaires). Each study, in turn, can have so-called collaborators managing the study's content. A collaborator is a user with additional (role-based) permissions on specific studies. Corona Check only contains one study, containing the questionnaire given in Table 3.11. All users are associated with this one study.

To collect data in a structured way, one study has one or more questionnaires in the entity of the same name. Each questionnaire can be versioned. The Corona Check study contains one questionnaire, available in each supported language. Questionnaires are structured with polymorphic building blocks called `Questionnaire Elements`. Elements can be of the type `Page Element`, `Text Element`, `Question Element`, `Headline Element` or `Media Element`. These building blocks can be used to create complex medi-

cal or psychological questionnaires meeting requirements for a variety of studies. The submitted answers for a questionnaire are serialized and stored in JSON in the entity `Answer Sheets` ($n = 86,912$). The entity also stores sensor data (e.g., location) and client device information (e.g., operating system) in the same table.

In addition, Corona Check provides tips (see section 3.3.3.3). Tips can also be managed and stored in different languages. A *like* feature allows users to rate the provided Corona tips. Tips were rated by 970 unique users. In order to give the user feedback immediately after submission of a questionnaire, such a questionnaire may reference to one or more key-rule pairs that are stored in the entity `Feedback`. Rules, evaluated on the client side, can be managed and adjusted dynamically.

3.3.3.5 | Medical Device Regulation

Recently, the requirements specifically for mHealth-related (mobile health) apps have been increasing. When we released Corona Check, we had to comply with the medical device regulation (MDR). With the MDR, strict rules have to be met regarding the validation and documentation of each software module. Especially in situations like the beginning of the coronavirus pandemic, when software solutions were required in as short a time as possible, such requirements can pose a risk for timely app releases. For more details refer to our publications related to the topic [252; 253; 260]. We were among the few apps that adhered to the MDR in the beginning of the COVID-19 pandemic with our mHealth system Corona Check. On top of the medical device regulation, the app stores of Google and Apple were especially cautious about allowing apps related to COVID-19 into their stores, likely to prevent allowing malicious apps. Overall, when releasing Corona Check, complying with all necessary regulations took time before the app could be found by the average user. Additionally, to all the regulations mentioned above, this study was approved by the ethics committee of the University of Würzburg with ethical approval no. 71/20-me on April 4, 2020.

3.3.4 | Results

To show that our mHealth system is a feasible solution that is able to reveal meaningful results, we present selected results using descriptive statistics. Therefore, we analyzed the users' self-reported data. Overall, at the time of data extraction, from all 145,223 users in the database, 52,267 (36%) filled out at least one assessment. From those 52,267 users, there were 86,912 assessments. To obtain the final dataset, we filtered the data set twice. First, we removed all assessments without a research release. Users could explicitly state if they agree to their data being used for research purposes (see question

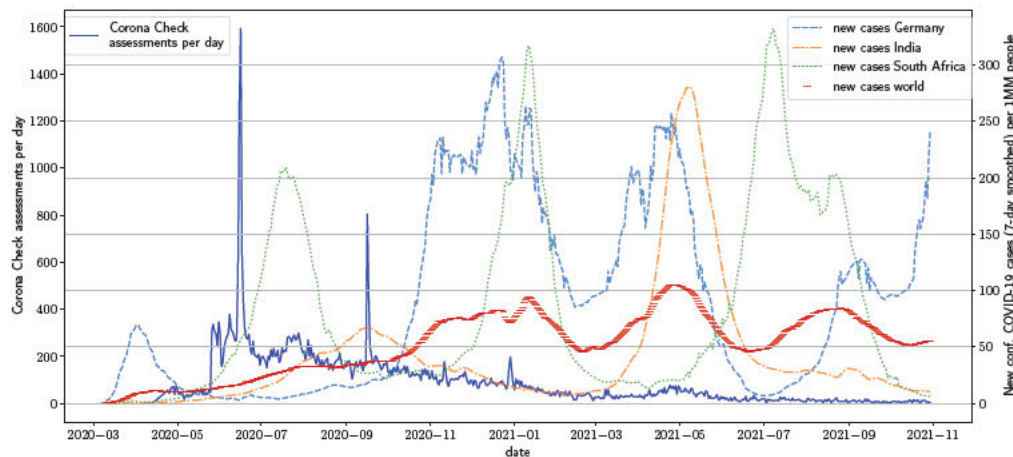


Figure 3.17: Corona Check assessments over time. For comparison, we also show new COVID-19 cases (7-day smoothed) per 1 million people.

7 in Table 3.11). This leaves us with 56,655 remaining assessments (65.2%). Second, we removed implausible assessments as follows: Each time, the questionnaire is filled out, the user enters age range and sex (questions 3 and 4). We classified an assessment as *implausible* if a user repeatedly filled out the questionnaire for himself/herself, but either the age or the sex differed from the first time he/she filled out the questionnaire.

The final dataset had 51,323 remaining assessments from 35,118 users, stemming from a total of 140 countries. Most assessments were filled out for the user himself/herself; 4,741 users (13.5% of all users) filled out 5,877 assessments (11.5% of all assessments) for others. For 36,212 assessments (70.6%), the users shared location information with us. Overall, 47,066 assessments (91.7%) were conducted with an Android device, the rest with an iOS device. Looking at multiple assessments, we observed that 80% of users filled out only one assessment. The remaining 20% (7,075 users) filled out 3.29 assessments on average ($SD=9.58$) and the mean time delta between first and last assessment was 18.88 days ($SD=55.43$).

Figure 3.17 shows the number of all assessments over time. There were more assessments in the earlier phases of the COVID-19 pandemic. The plot shows two peaks for June 2020 and September 2020. We added the number of confirmed new cases for Germany, India, South Africa, and the world in the plot [261]². In Table 3.12, we present details about the demographic information of the included users.

We present the global completion behavior of assessments in Fig. 3.18. It shows the number of assessments completed per country.

²Via <https://github.com/owid/covid-19-data/tree/master/public/data>; accessed 2022-01-11

Sex Age	00-09	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80+
female	2.4%	22.5%	25.1%	19.2%	13.3%	8.6%	5.9%	2.2%	0.7%
male	1.9%	19.3%	26.3%	17.3%	10.7%	8.6%	8.2%	5.7%	1.9%
diverse	11.9%	16.1%	22.5%	10.1%	9.6%	6.0%	5.5%	3.7%	14.7%
no answer	9.6%	31.7%	21.0%	11.4%	5.9%	3.1%	2.6%	1.8%	12.9%
all	2.2%	20.6%	25.8%	17.9%	11.6%	8.5%	7.4%	4.5%	1.7%

Table 3.12: Age distribution grouped by sex for assessments completed as of October 30, 2021. The percentages sum up to 100% line by line. The age group of 20-29 is the most common, with 12,957 assessments, followed by age groups 10-19 (10,334), and 30-39 (8,977).

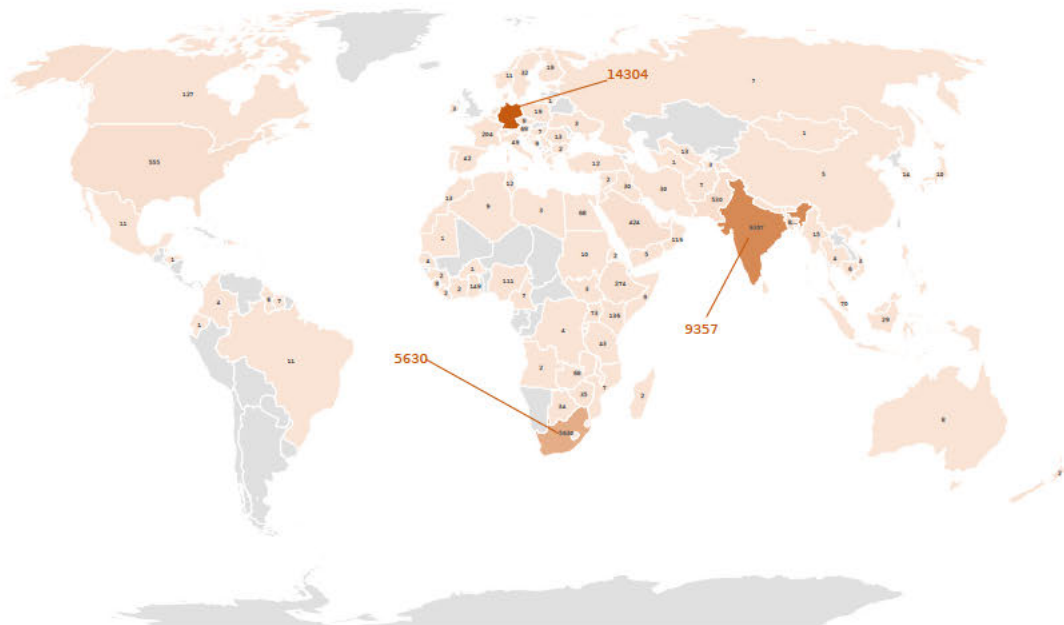


Figure 3.18: Number of completed assessments per country as of October 30, 2021. A total of 51,323 assessments from 140 countries have been completed and was available for research. The top 3 countries were Germany (14,304, 27.9%), India (9,357, 18.2%), and South Africa (5,630, 11%). 125 countries were represented with less than 100 assessments, 71 countries with less than 10 assessments.

The distribution of the reported symptoms between groups did not differ significantly when stratified for age group and sex (Table 3.13). We also looked at the distribution of symptoms between countries. We only investigated countries with at least 51 users. An ANOVA test could not detect statistically significant differences in symptom distributions between countries. However, the number of reported symptoms differed between countries. Users from India reported more than twice as many symptoms per assessment as those from Germany (2.81 vs. 1.27). Users in France reported the fewest symptoms

Table 3.13: Age by sex and symptoms. Each line adds up to 100%. An ANOVA test did not reveal any differences between the symptom distributions per group.

Age	Sex	n	Fever	Sore throat	Runny nose	Cough	Loss smell	Loss taste	Shortness breath	Headache	Muscle pain	Diarrhea	General weakness
00-09	female	1119	14.9%	9.4%	13.1%	14.6%	6.4%	5.7%	7.2%	11.0%	8.0%	0.4%	9.1%
	male	1610	16.1%	9.2%	12.9%	14.8%	6.2%	6.3%	7.8%	9.7%	8.2%	0.4%	8.3%
	diverse	109	15.6%	9.2%	10.1%	10.1%	10.1%	8.3%	11.0%	9.2%	8.3%	0.9%	7.3%
	no answer	152	14.5%	7.9%	9.2%	13.2%	7.2%	6.6%	7.9%	11.8%	9.9%	2.0%	9.9%
	all	2990	15.6%	9.2%	12.7%	14.5%	6.5%	6.2%	7.7%	10.3%	8.2%	0.5%	8.6%
10-19	female	9038	9.3%	10.3%	11.6%	13.6%	5.9%	5.1%	7.4%	16.1%	9.1%	0.3%	11.5%
	male	15421	12.4%	9.2%	10.1%	14.5%	6.6%	5.9%	8.0%	12.6%	9.2%	0.2%	11.3%
	diverse	136	8.1%	11.8%	8.1%	14.0%	8.1%	7.4%	9.6%	13.2%	10.3%	0.0%	9.6%
	no answer	430	13.3%	8.6%	10.0%	14.9%	7.7%	6.5%	9.3%	11.6%	8.8%	0.0%	9.3%
	all	25025	11.3%	9.6%	10.6%	14.2%	6.4%	5.6%	7.8%	13.9%	9.1%	0.2%	11.3%
20-29	female	10520	9.2%	10.4%	10.2%	12.6%	6.3%	6.2%	7.0%	15.9%	10.1%	0.3%	11.9%
	male	21658	13.7%	9.0%	9.2%	12.1%	7.3%	7.0%	7.9%	11.6%	9.8%	0.2%	12.2%
	diverse	161	12.4%	11.8%	8.1%	11.2%	6.8%	7.5%	8.7%	11.2%	9.3%	1.2%	11.8%
	no answer	386	14.2%	9.6%	9.6%	10.9%	8.5%	7.8%	9.3%	9.8%	9.3%	0.0%	10.9%
	all	32725	12.3%	9.4%	9.5%	12.3%	7.0%	6.7%	7.6%	13.0%	9.9%	0.2%	12.1%
30-39	female	7571	8.0%	11.2%	11.1%	12.5%	5.2%	5.5%	6.9%	15.9%	11.5%	0.4%	11.8%
	male	13005	11.3%	9.6%	9.1%	12.9%	6.8%	6.5%	7.5%	12.4%	11.4%	0.4%	12.3%
	diverse	66	15.2%	10.6%	9.1%	9.1%	9.1%	7.6%	7.6%	12.1%	9.1%	0.0%	10.6%
	no answer	186	16.1%	9.7%	10.2%	11.3%	7.5%	7.0%	8.6%	9.7%	8.1%	0.0%	11.8%
	all	20828	10.2%	10.2%	9.9%	12.7%	6.2%	6.1%	7.3%	13.6%	11.4%	0.4%	12.1%
40-49	female	4217	7.8%	10.4%	10.3%	12.7%	4.5%	4.7%	6.4%	17.1%	12.7%	0.4%	13.0%
	male	6439	10.4%	9.4%	9.7%	13.9%	5.2%	5.2%	7.4%	13.4%	11.4%	0.5%	13.5%
	diverse	97	13.4%	11.3%	9.3%	12.4%	8.2%	8.2%	8.2%	11.3%	9.3%	0.0%	8.2%
	no answer	96	9.4%	9.4%	8.3%	11.5%	9.4%	10.4%	8.3%	10.4%	8.3%	0.0%	14.6%
	all	10849	9.4%	9.8%	9.9%	13.4%	5.0%	5.1%	7.0%	14.8%	11.9%	0.4%	13.3%
50-59	female	2378	6.5%	10.4%	11.3%	13.8%	3.7%	4.5%	7.5%	15.8%	13.4%	0.6%	12.6%
	male	3265	7.8%	8.5%	11.0%	15.1%	3.4%	3.7%	8.3%	12.4%	14.3%	0.9%	14.6%
	diverse	54	13.0%	9.3%	9.3%	11.1%	9.3%	7.4%	9.3%	13.0%	9.3%	1.9%	7.4%
	no answer	18	33.3%	16.7%	5.6%	5.6%	0.0%	0.0%	5.6%	11.1%	5.6%	0.0%	16.7%
	all	5715	7.4%	9.4%	11.1%	14.5%	3.6%	4.0%	8.0%	13.8%	13.8%	0.8%	13.7%
60-69	female	1300	5.9%	6.7%	7.4%	12.8%	2.5%	3.3%	6.9%	23.2%	21.2%	0.5%	9.5%
	male	1708	7.4%	7.0%	11.8%	15.0%	4.0%	3.6%	10.0%	11.1%	15.2%	1.1%	13.8%
	diverse	55	12.7%	10.9%	9.1%	12.7%	9.1%	9.1%	10.9%	9.1%	9.1%	0.0%	7.3%
	no answer	40	10.0%	10.0%	15.0%	5.0%	7.5%	7.5%	7.5%	10.0%	17.5%	0.0%	10.0%
	all	3103	6.9%	7.0%	9.9%	13.9%	3.5%	3.6%	8.7%	16.1%	17.6%	0.8%	11.9%
70-79	female	348	9.2%	8.0%	10.1%	12.9%	3.4%	4.0%	12.1%	11.8%	15.2%	0.3%	12.9%
	male	1445	6.9%	6.4%	12.3%	12.4%	10.6%	10.2%	9.1%	8.6%	12.5%	0.7%	10.4%
	diverse	50	12.0%	8.0%	10.0%	8.0%	10.0%	10.0%	8.0%	12.0%	10.0%	0.0%	12.0%
	no answer	22	13.6%	9.1%	9.1%	13.6%	4.5%	4.5%	13.6%	13.6%	13.6%	0.0%	4.5%
	all	1865	7.5%	6.8%	11.8%	12.4%	9.2%	9.0%	9.7%	9.3%	13.0%	0.6%	10.8%
80+	female	411	13.1%	8.8%	8.8%	11.2%	8.5%	9.2%	10.0%	9.2%	11.2%	0.7%	9.2%
	male	1178	12.8%	8.8%	9.9%	10.5%	8.7%	8.2%	9.9%	9.8%	10.1%	1.1%	10.0%
	diverse	185	12.4%	8.6%	9.2%	10.3%	9.2%	10.3%	10.8%	9.7%	9.2%	0.5%	9.7%
	no answer	384	12.8%	9.9%	9.1%	9.9%	9.4%	9.1%	9.6%	9.6%	10.2%	0.8%	9.6%
	all	2158	12.8%	9.0%	9.5%	10.5%	8.9%	8.8%	10.0%	9.6%	10.2%	0.9%	9.8%

at 0.38 per assessment. Users from Uganda reported the most symptoms at 4.03 per assessment. The average number of reported symptoms per assessment per country is shown in Table 3.14. Although the completion of the questionnaire differed between countries, the distributions of symptoms seemed to be rather similar.

Corona Check overall contains 31 general tips on hygiene, see section 3.3.3.3. A total of 3,538 ratings were submitted with an average rating of 3.7 out of 5 stars (SD = 1.67). The top-rated tips were (in descending order): (i) protecting wounds, (ii) how to behave in daily life, (iii) when to wash hands, (iv) masks that cover the nose and mouth, and (v)

Country	O*	fever	sorethroat	runnynose	cough	losssmell	losttaste	shortnessbreath	bradace	musclepain	diarhea	generalweakness
Arab Emirates	2.63	12.9%	9.7%	10.2%	13.1%	7.1%	6.8%	8.1%	12.3%	8.7%	0.8%	9.7%
Austria	1.57	5.6%	8.3%	16.7%	10.2%	6.5%	0.9%	7.4%	22.2%	10.2%	0.9%	11.1%
Bangladesh	3.15	16.5%	7.9%	7.9%	11.5%	6.6%	6.5%	8.1%	10.8%	11.0%	0.5%	12.6%
Belgium	1.28	1.3%	16.9%	14.3%	19.5%	1.3%	2.6%	11.7%	10.4%	6.5%	5.2%	10.4%
Canada	0.45	5.3%	10.5%	15.8%	10.5%	7.0%	5.3%	5.3%	12.3%	19.3%	5.3%	3.5%
Switzerland	1.32	5.3%	12.0%	17.3%	13.3%	2.7%	2.7%	6.7%	15.3%	8.0%	0.7%	16.0%
Germany	1.27	4.5%	10.7%	13.7%	14.5%	3.7%	3.6%	7.6%	16.2%	12.5%	1.0%	11.9%
Egypt	4.01	7.7%	11.4%	10.6%	11.0%	6.6%	7.0%	7.0%	16.1%	10.6%	0.4%	11.7%
Ethiopia	2.86	15.3%	7.5%	8.3%	12.0%	8.2%	7.9%	8.8%	10.8%	9.9%	0.4%	11.0%
France	0.38	5.2%	7.8%	15.6%	18.2%	1.3%	2.6%	10.4%	14.3%	18.2%	1.3%	5.2%
United Kingdom	1.53	11.0%	12.6%	11.5%	19.2%	6.0%	6.6%	5.5%	10.4%	7.7%	0.0%	9.3%
Ghana	2.74	9.8%	8.6%	10.8%	13.5%	5.9%	4.9%	6.4%	19.1%	9.3%	0.2%	11.5%
India	2.81	14.1%	8.3%	8.1%	12.7%	7.3%	7.1%	7.9%	11.3%	9.9%	0.1%	13.1%
Kenya	2.83	9.1%	8.1%	9.6%	11.2%	7.5%	8.3%	6.8%	17.1%	13.0%	0.5%	8.8%
Sri Lanka	3.43	11.9%	9.2%	6.9%	11.9%	7.2%	7.2%	10.6%	14.7%	11.1%	0.8%	8.3%
Malaysia	3.31	15.1%	7.3%	8.6%	11.6%	7.8%	7.8%	8.2%	10.3%	12.1%	0.4%	10.8%
Nigeria	2.22	8.9%	12.2%	6.5%	11.4%	8.1%	6.9%	4.5%	16.3%	12.6%	0.0%	12.6%
Netherlands	1.39	8.0%	13.0%	12.2%	15.9%	4.2%	3.6%	8.7%	16.1%	8.9%	2.4%	7.1%
Nepal	3.31	12.9%	8.4%	8.7%	10.8%	6.8%	6.4%	8.2%	13.5%	10.2%	0.5%	13.5%
Oman	2.97	14.2%	9.0%	8.4%	11.6%	6.7%	7.8%	8.7%	11.6%	11.3%	0.3%	10.2%
Philippines	2.31	5.8%	8.2%	9.9%	16.4%	8.8%	7.6%	9.4%	14.6%	12.9%	0.6%	5.8%
Pakistan	3.72	12.4%	9.2%	8.8%	11.9%	6.7%	6.9%	8.8%	10.9%	10.4%	0.3%	13.6%
Qatar	1.93	14.5%	10.0%	6.4%	7.3%	6.4%	6.4%	10.0%	14.5%	11.8%	3.6%	9.1%
Saudi Arabia	2.96	10.2%	7.5%	9.1%	10.4%	7.4%	7.2%	7.4%	11.0%	10.8%	1.0%	10.2%
Uganda	4.03	11.6%	10.5%	9.9%	13.9%	7.8%	5.4%	7.8%	12.6%	8.8%	0.0%	11.6%
USA	1.17	7.7%	11.6%	11.1%	14.3%	4.6%	4.8%	8.8%	14.8%	11.6%	0.0%	10.8%
South Africa	1.60	8.6%	10.6%	10.9%	14.1%	6.6%	6.3%	6.3%	16.5%	10.7%	0.0%	9.4%
Zambia	2.84	9.8%	8.8%	9.3%	14.5%	8.3%	5.2%	7.8%	14.0%	10.4%	0.5%	11.4%

Table 3.14: Distribution of reported symptoms of assessments stratified by country. The analysis is only applied to countries represented by at least 51 users. *Average number of reported symptoms per assessment.

handling surfaces and objects. The following tips received the lowest ratings: (i) How to cover sneezing or coughing, (ii) tissues, (iii) smear infection, (iv) shaking hands and hugging, and (v) traveling. An overview of all tips and the distribution of their ratings is given in Fig. 3.19. We further analyzed who rated the tips. Table 3.15 shows the age distribution for all users compared to those who rated at least one of the tips. We found that older users rated more often than younger users. While the most common age group for all users was 20-29 years, the one for users who rated the tips was 60-69. Table 3.16 shows the location distribution of all users compared to those who rated the tips. The

Age	Users who rated	All users
00-09	2.1%	2.2%
10-19	10.7%	20.6%
20-29	8.6%	25.8%
30-39	8.8%	17.9%
40-49	13.0%	11.6%
50-59	13.4%	8.5%
60-69	26.6%	7.4%
70-79	14.3%	4.5%
80+	2.5%	1.7%

Table 3.15: Age distributions of users that rated compared to all users.

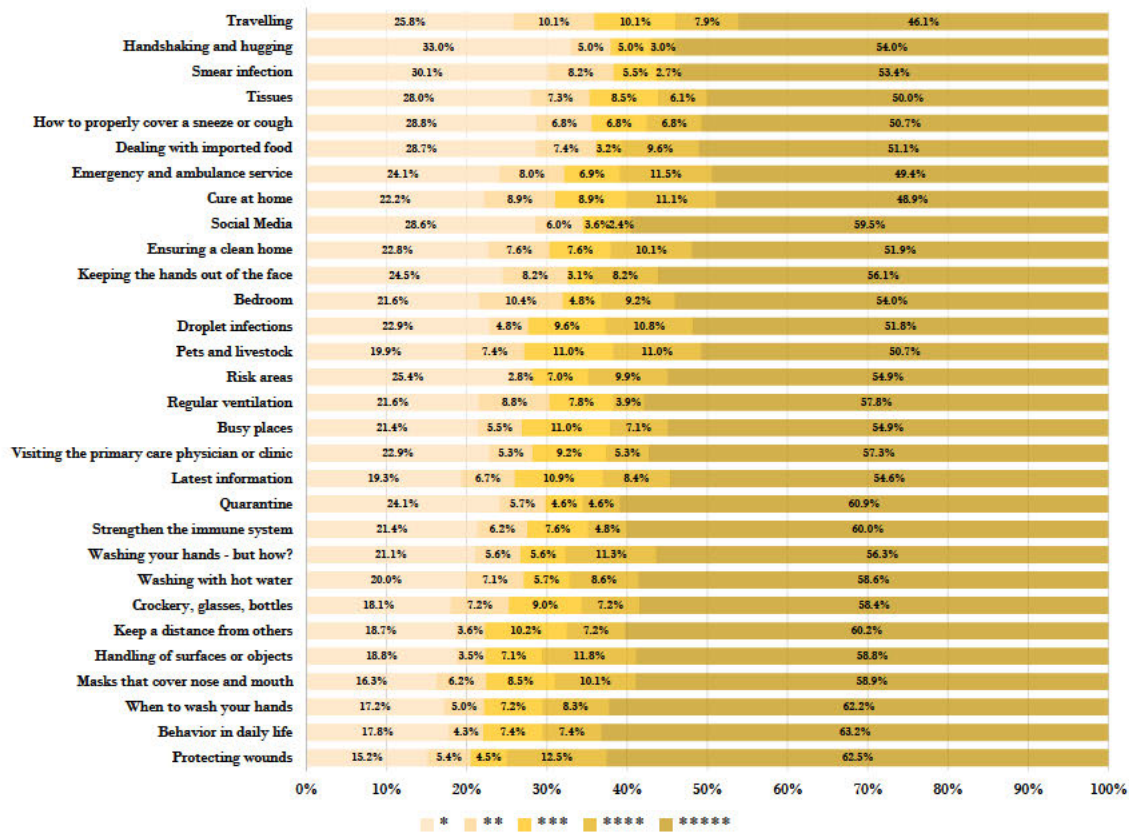


Figure 3.19: Distribution of star ratings of the Corona Check tips, which were available as general information on hygiene. Each tip was rated between 71 and 250 times, with an overall average rating of 3.7 out of 5 stars. In the first rows are the least popular tips, in the last the most popular ones.

results show that users in Germany were largely overrepresented in the group of users who rated the tips.

Country	Users who rated	All users
Germany	66.2%	27.9%
India	6.5%	18.2%
Location n. A.	16.6%	29.4%
South Africa	3.7%	11.0%

Table 3.16: Location distribution of users who rated compared to all users. Only the top 4 locations are shown. Percentages refer to the whole dataset.

3.3.5 | Limitations

Several limitations of Corona Check have been revealed during its practical use. We quickly encountered the need to adapt the questionnaire and its feedback texts in frequent cycles as an important feature due to changing recommendations and new findings regarding COVID-19-related symptoms. However, providing a robust mechanism that does not confuse or distract users with respect to different questionnaire versions and frequent changing feedback texts is important, but not simple. As any change to the app must be considered in the context of the medical device regulation (and time matters during COVID-19), we decided to show only the most recent version of the questionnaire as well as the most recent feedback texts, which mitigated measures for the medical device regulation and saved us time. Although we had not complaints about our approach, possibly, a more fine-grained approach might fit the users' needs and the phases of the pandemic better.

In addition to the provided questionnaire, we quickly saw the need to display further information in a smart way when filling out the questionnaire. For example, by the time certain regions have been declared risk regions, it was important to update this information as well as letting travelers know about possible consequences when traveling to or returning from these regions. However, the provision of the information was necessary in a way that the existing questionnaire-procedure can be distinguished from this new information.

Another limitation that we encountered was that for users who filled out questionnaires multiple times, it had to be checked, whether all completed questionnaires can be used for evaluation or if the users just wanted to 'play' with all the combinations of filling out the questionnaire to see what feedback is possible. The gambling behavior, in turn, might affect the validity of the data. Corona Check was provided in German or English. More languages could increase its use in more countries and multilingual societies and thus increase the number of filled-out questionnaires. This may be helpful for better insights into the development of the pandemic in a rather short period of time.

3.3.6 | Discussion

To the best of our knowledge, we are the first to report about a large-scale deployment of a mHealth system for assessing potential COVID-19 symptoms. Corona Check provided specific symptom-related advice as well as general tips for behavior and hygiene. We highlighted the technical details of Corona Check and analyzed the collected data.

First of all, we note that Corona Check was not as widely known as some contact tracing apps, which received adoption rates as high as 50% of the population (Germany) [233]. We did not advertise extensively for Corona Check, and self-assessment apps did not receive as much media coverage as contact tracing apps.

We found that only 36% of the users filled out at least one assessment. Out of these, 80% only filled out one. There could be several reasons for this. Maybe the news and tips sufficed for many users' purposes. Maybe users just wanted to see what the app does and then decided to not use it or to use it only once. Overall, we found that more younger users used our app (see Table 3.12); most users are below 40 years of age. This is in line with the idea that younger users tend to be more tech-savvy and more likely to use an app instead of calling a hotline. For 65.2% of all Corona Check assessments, the users agreed to their data being used for research purposes. In the final dataset, we had geolocation information for 70.6% of the assessments. Thus, in line with some of our previous studies, we found that most users are willing to share their data with researchers [262].

We did not observe that the number of new confirmed cases influenced the number of Corona Check assessments (see Figure 3.17). Likely, the two peaks in Corona Check assessments in June 2020 and September 2020 are due to some news or social media posts creating a brief period of increased public interest in Corona Check. A broader active advertisement of mHealth systems like Corona Check might create a larger user base. Then, we would expect to see some correlation between in-app-assessments and new cases. After the two peaks, we observed a steady decline in the number of assessments per day. There could be two reasons for that. First, existing users might lose interest in the app and fewer new users were on-boarding. Second, with passing time, public knowledge about corona increased, as well as the availability of testing stations, minimizing the need for an app like Corona Check. Thus, our user data strongly supports the notion that an app-based mHealth system for the population is particularly important in the early stage of a pandemic. Note that testing was not widely available during the beginning of the pandemic and Corona Check did not offer to register test results. Hence, we were not able to investigate to what extent the high-risk warning indicated a real infection.

Regarding the symptoms entered in Corona Check, we did not find statistically significant differences between age groups, or between countries. This may indicate that symptoms were independent of these variables. Regarding the general hygiene tips in Corona Check, overall, the ratings indicated that they were perceived as helpful, see Fig. 3.19. We observed that the proportion of older users rating the tips was higher than the proportion of younger users (see Table 3.15). Raters from Germany were dispropor-

tionately overrepresented among the raters of the tips (see Table 3.16). We presume that users in Germany might have been aware that Corona Check was made in Germany, leading to higher identification with the app or trust in the app, and thus, a prolonged usage including rating the tips.

Overall, we have shown that an mHealth system such as Corona Check can help support much of the functionality that a telephone hotline by, e.g., authorities or health insurances, would serve. With increasing public knowledge about symptoms related to the new virus and broadly available testing stations, the need for an mHealth system for detecting coronavirus infections might be reduced. Thus, especially during the early phase of the pandemic, Corona Check was a valuable contribution in fighting the global COVID-19 pandemic.

3.4 | How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare.

- **Authors** | Allgaier, Johannes; Mulansky, Lena; Draelos, Rachel and Pryss, Rüdiger
- **Published in** | Artificial Intelligence in Medicine, 143, 102616, 2023.
- **Available at** | <https://www.sciencedirect.com/science/article/pii/S0933365723001306>

Abstract

Background Medical use cases for machine learning (ML) are growing exponentially. The first hospitals are already using ML systems as decision support systems in their daily routine. At the same time, most ML systems are still opaque and it is not clear how these systems arrive at their predictions.

Methods In this paper, we provide a brief overview of the taxonomy of explainability methods and review popular methods. In addition, we conduct a systematic literature search on PubMed to investigate which explainable artificial intelligence (XAI) methods are used in 450 specific medical supervised ML use cases, how the use of XAI methods has emerged recently, and how the precision of describing ML pipelines has evolved over the past 20 years.

Results A large fraction of publications with ML use cases do not use XAI methods at all to explain ML predictions. However, when XAI methods are used, open-source and model-agnostic explanation methods are more commonly used, with SHapley Additive exPlanations (SHAP) and Gradient Class Activation Mapping (Grad-CAM) for tabular and image data leading the way. ML pipelines have been described in increasing detail and uniformity in recent years. However, the willingness to share data and code has stagnated at about one-quarter.

Conclusions XAI methods are mainly used when their application requires little effort. The homogenization of reports in ML use cases facilitates the comparability of work and should be advanced in the coming years. Experts who can mediate between the worlds of informatics and medicine will become more and more in demand when using ML systems due to the high complexity of the domain.

3.4.1 | Introduction

Artificial Intelligence (AI) in healthcare holds many opportunities and risks and has attracted great public interest. To date, however, experts involved in the development of Machine Learning (ML) systems come from diverse backgrounds, and the gap between ML engineers and healthcare providers, and often also other researchers, is wide. Methods that explain the predictions of complex algorithms in a user-friendly way can increase adoption and trust [263]. The use of ML systems for appropriate medical use cases has the potential to reduce costs, save time, increase treatment quality, and improve patient care.

ML systems can be categorized based on whether they can *replace* or *supplement* a healthcare provider. To date, there are no ML systems capable of replacing a healthcare provider; to the best of our knowledge, we did not find any system that appears to be sufficiently powerful or interpretable to operate safely without human supervision. In a few cases, ML systems are currently being used to *supplement* health care. Some of these are listed in an online database [264]. However, the companies using these AI systems in hospitals do not provide detailed information on their websites about whether these systems include explainability methods.

The medical specialties with the most ML activity are radiology and pathology, as they are both image-based and therefore ideally suited to recent advances in computer vision techniques [265]. ML systems applied to radiology images have the potential to reduce radiologist error rates from 3 - 5 % [266] by alerting radiologists to potentially missed diagnoses, extend specialist expertise to under-supplied regions, where only one radiologist may be available for millions of patients [267], or improve triage by bringing scans with potentially urgent findings to the top of the physician's queue for earlier interpretation. In pathology, AI systems can speed up the interpretation of large slides by automatically identifying the most important areas for the pathologist to examine [268]. There is also interest in developing AI systems for dermatology [269], cardiology [270], genetics [271], intensive care [272], oncology [273], and gastroenterology [274]. In the future, ML systems focused on augmentation may influence administrative or research activities, such as chart review [275], in addition to clinical care. So, we see that there are many approaches and good reasons to implement AI systems in the health care context. But is it also necessary to make AI systems explainable in this context? And if so, explainable to whom?

Is it necessary to make ML models explainable in medicine? Explainability of AI systems is not always necessary, or if the benefits outweigh the costs of explainability

too much, then perhaps it can be dispensed with. For example, in logistics, if a package is occasionally misclassified, and therefore sent somewhere else, this need not be a major problem. However, the situation is different for decisions involving the health of patients. Explainability is therefore crucial for medical ML systems and benefits all parties involved: patients, physicians, governments, ML engineers, and other decision makers in the healthcare system. All these parties have a legitimate interest in fair, unbiased, reliable, and reasonable AI based on medical properties rather than spurious correlations [276]. Transparent AI that provides explanations for its predictions facilitates these goals by enabling users to better understand the factors that contributed to a prediction. Explanation methods also enable governments to more effectively regulate AI systems through audits, and machine learning engineers to more easily maintain and improve their models. Stakeholders expect decision support systems to be transparent and to fit seamlessly into existing workflows [263; 277]. Above all, however, transparency includes being explainable.

Unfortunately, explainability methods are underutilized in medical ML research. It is already an immense amount of work to define a medical problem suitable for a ML solution, obtain the necessary data, clean the data so that it can be used for modeling, develop a model, and refine the model to achieve high performance. Therefore, once one is past this hurdle, in many cases the inclusion of explainability is no longer considered or was not planned for in the first place.

With this work, we hope to facilitate the incorporation of explainability methods into medical ML through two main contributions. First, we provide a representative overview of the major classes of ML interpretability methods and highlight the advantages and limitations of the various approaches. Second, we analyze the extent to which previously published papers in medical ML use explainability methods and quantify which methods are used and how they are presented. We hope that this will allow researchers who have not previously been familiar with explainable ML to select a method or class of methods that are appropriate for their area of research and to integrate explainability in the future. To achieve our goal, we conducted a comprehensive and systematic literature search. Inspired by recent analyses (e.g., [31]) and several discussions with medical professionals, we systematically searched PubMed using PRISMA guidelines. We have paid particular attention to the following aspects, the combination of which has received little attention to date:

- Is there a concrete medical supervised ML use case that uses interpretability methods?
- Who is potentially able to understand the XAI methods explained in the paper?

- Which kind of data is used, and how well is ML pipeline described?
- Do authors provide their source code and data?

With these aspects in mind, our PubMed search found 2,568 papers, of which 450 remained after applying exclusion criteria. In the following, we present our approach and show that the field is changing dynamically.

3.4.2 | Related Work

This section is divided into three subsections. The first section deals with other work on explainability methods and the description of the XAI taxonomy. The second section deals with reviews from similar or related XAI areas that follow slightly different naming conventions. The third section deals with cutting edge topics such as causal ML. Although the topic of XAI is still young, even in medicine, we consider this subdivision already important and discuss the related works along the categories.

There are other reviews of explainability methods, i. e., Ward prepared a summary table of explainability methods, available [here](#). Although this paper mentions essential XAI methods, the methods are not classified according to their application to tabular data or image data. In addition, some more recent methods are missing. [This GitHub repository](#) contains a large Markdown table with hyperlinks to source code for explainability methods organized by year. However, this repository is mainly focused on image classification rather than medical data. Tjoa and Guan provide an overview of some interpretability methods related to medicine. However, they follow a different taxonomy for ML interpretability methods that we have not found in other works, which makes comparability and classification difficult [278]. Another systematic review considers XAI systems in the medical field [279]. The authors found that post-hoc methods were more common than intrinsic methods in the papers reviewed, and they discuss human-in-the-loop and inclusion of domain experts³. However, they did not examine why other papers did not use XAI. Linardatos et al. [19] published a comprehensive collection of existing methods of ML interpretation methods. They propose an alternative taxonomy to allow a multi-perspective comparison between techniques. Methods are categorized into four main groups by intended use: *methods for explaining complex black-box models*, *methods for building white-box models*, *methods for limiting discrimination and improving fairness in models*, and *methods for analyzing the sensitivity of model predictions*. In the group of methods that explain black-box models, the authors further distinguish between

³By domain experts, we mean experts in the field to which the ML algorithm is applied. For example, in healthcare use cases, this could be physicians.

black box deep learning models and arbitrary black box models. The second category contains methods that create easy-to-understand models, while the third class includes techniques that focus exclusively on the discrimination, inequality, and impartiality of an ML algorithm and evaluate it with respect to these properties. The methods in the last group are applied to evaluate ML algorithms in terms of reliability and sensitivity to ensure that their predictions are credible and consistent [19]. Linardatos et al. do not address the medical application of the XAI methods presented, nor do they discuss real-world use cases of the approaches. In the recent papers [280], [281], [282], [283], and [284], the authors focus on XAI in a medical context. However, they each consider a specific medical subspecialty rather than a general view of medicine. In these works, the authors also did not analyze whether source code is provided, what stakeholders benefit from the XAI, and for which data format the presented methods are suitable.

There are also reviews with different wordings, i. e., Antoniadi et al. performed a systematic literature review for clinical decision support systems (CDSS). The main finding was the absence of XAI in CDSS for tabular and image data.[285]. Quinn et al. provide an overview of the current state of machine learning in healthcare and provide an optimistic and pessimistic scenario for future diagnosis of AI systems in healthcare. However, they do not go into detail about current explanatory methods, but rather trace the historical development of ML in healthcare [286]. Other related work in the area of XAI research include i. e. Holzinger et. al., who argue beyond explainability by saying that the domain expert understands an ML system better because s/he knows the causality of the relationships, but the system only knows the data [287]. In this context, Holzinger et al. emphasize the importance of causality relations in XAI, but also mention that so far these cannot be given by the algorithm but require domain knowledge. Adida and Berrada provide an overview of XAI in general and classify common ML interpretation methods using the taxonomy explained in the Methods section of this paper. However, they do not do so in terms of medicine, nor do they rank the methods in terms of their applicability to tabular data or neural networks [22]. Longo et al. address the challenges and emphasize the relevance of XAI in sensitive sectors such as medicine or law, but the focus is on XAI in general rather than medicine in particular [288].

While existing work has mostly examined the existing literature for XAI applications in specific medical sub-specialties, to our knowledge, there has been no general literature review of a similar scope to our work for XAI applications in the medical field overall.

3.4.3 | Explainability Methods

By the word *model*, we generally mean the model learned by the system after performing some learning algorithm. Some machine learning models are inherently explainable, including linear regression, logistic regression, generalized linear models, or decision trees. Other models, such as neural networks, are black box by default, but can be augmented with explainability methods. These “add-on” explainability methods for otherwise non-interpretable models are the focus of this section. For a more detailed overview of the explainability methods we consider, see our supplementary material at [GitHub](#). Explainability is also referred to in the literature as interpretability, intelligibility [289; 290], causability [287], or understandability [21]. There is a tendency for “explainability” to refer to model-specific methods and “interpretability” to refer to inherently interpretable ML models or model-agnostic methods, but there is no consensus in the research community. We thus use the terms explainability and interpretability interchangeably in this paper, under the following definition:

Explainability method (synonym *interpretability method*): A method that enables humans to understand why a model makes certain predictions.

3.4.3.1 | Trustworthy vs. untrustworthy explainability methods

Some explainability methods are trustworthy, meaning that their explanations are provably guaranteed to reflect the model’s computations. A trustworthy explainability method can be used to assess how a model performs and help distinguish between performing and non-performing models. Models that do not perform should be tested further, while models that do perform could be moved into a deployment process. Whether a model performs or not depends on the use case and the chosen metrics to optimize, such as Mean Squared Error for regression tasks or weighted F1-scores for multi-class classification tasks. When a trustworthy explainability method is applied to a **non-performing model**, the explanations may seem strange or unexpected - for example, a neural network that uses metal tokens and postprocessing artefacts to predict pneumonia from chest x-rays [276] or a neural network that highlights snow to explain its classification of a *wolf*. The key is that the explanations can be used to conclude that the model is non-performing, because the explanation method is trustworthy. When a trustworthy explainability method is applied to a **performing model**, then the explanations will make sense even under maximum scrutiny by a human domain expert. In the ideal case, the explanation of a performing model will match the explanation of a group of domain experts, and the performing model could then be considered for

deployment in a real-world setting. Some explainability methods are not trustworthy because they do not come with mathematical guarantees that they reflect the model's computations [55]. For example, the explainability method Grad-CAM is popular and highly cited but has recently been shown to sometimes produce misleading explanations that do not represent how the model makes predictions [291]. Grad-CAM is thus not a trustworthy explainability method and cannot be used to draw conclusions about whether a model is non-performing or good. Because of the flaw in Grad-CAM, a new method called HiResCAM [291] was developed which does come with mathematical guarantees that it accurately reflects the underlying model, and thus HiResCAM is a trustworthy explainability method.

Is there a difference between the terms explainability and interpretability? We review some definitions of interpretability and explainability in the literature. Sometimes the definitions agree, and sometimes they contradict each other. We use these terms interchangeably. Interpretability answers the question of how the models work, while explainability answers the question of what else the model says according to [21]. [292] defines interpretability as the ability to explain to a human in terms that can be understood. [55] in turn states that post hoc methods can be considered examples of explainability, while intrinsic methods can be considered examples of interpretability. [293] says that interpretability is the degree to which a human can understand the cause of a decision. Explainability in the technical sense highlights decision-relevant parts of the used representations of the algorithms and active parts in the algorithmic model that [...] contribute to model accuracy [...] [287]. Interpretability, in turn, is the extent to which a human can consistently predict the outcome of the model [106]. Gilpin et al. argue that explainable models are interpretable by default, but not vice versa. They describe explainability as models that can summarize the reasons for neural network behavior [294]. [Glassboxmedicine](#) states that interpretability means that the algorithm is intrinsically designed to establish a relationship between input and output that is understandable to humans, such as in linear regression, and counters that explainability means that the algorithm's decision making can be understood, even if it is abstractly detached from a human logic. For example, a deep-learning algorithm can explain a wolf by highlighting snow.

Explainability is not causality, algorithmic transparency, or simple input variables. Explainability methods typically specify which parts of the input contribute to the output of a model, but do not specify causal relationships or indicate *how* particular parts of the input affect a prediction. Explainability is also distinct from algorithmic transparency, which is a clear description of the algorithm's implementation and training process. In

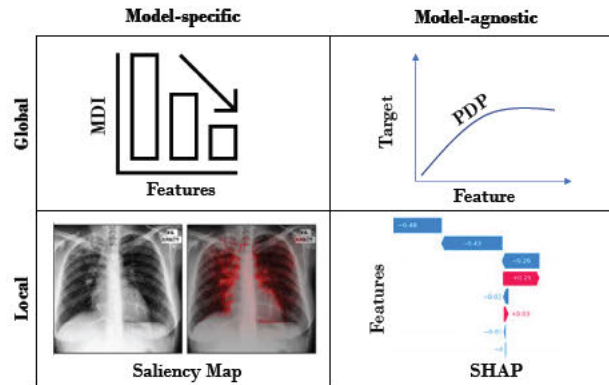


Figure 3.20: Taxonomy concept used to classify the XAI approaches, following [19; 20; 21; 22]. The methods outlined in the quadrants are exemplary representatives of this category. MDI = Mean Decrease Impurity, PDP = Partial Dependence Plots, SHAP = Shapley Additive Explanations.

our opinion, understanding what an input variable means (e.g., the dictionary definition of the variable "age") does not make the model explainable.

3.4.3.2 | Types of explainability methods

Explainability methods can be classified as model-agnostic or model-specific and as global or local, following the taxonomy of [19; 20; 21; 22], and as shown in Figure 3.20. Model-agnostic methods are independent of the structure of the ML algorithm, while model-specific methods can only be applied to specific classes of models. Global methods explain the model as a whole, while local methods explain a single prediction. In the following subsections, representative examples of different classes of interpretability methods are described. Following our filter criteria of our literature review, the following described machine learning methods have the purpose to primarily explain supervised machine learning methods. More technical explanations are aimed at ML engineers, context dependent methods like saliency maps are rather aimed at domain experts, simpler methods like partial dependence plots or decision trees are aimed primarily but not exclusively at non-ML experts and non-domain experts like, i. e., patients.

3.4.3.3 | Global Model-Agnostic Methods

Global, model-agnostic methods describe the overall average behavior of models and are applicable to a wide range of machine learning models. Typical examples of methods in this category are Partial Dependence Plots (PDPs), Permutation Feature Importance,

Leave One Covariate Out (LOCO) and Maximum Mean Discrepancy Critic (MMD-Critic), and Permutation Importance.

A partial dependence plot (PDP) illustrates the marginal effect of a feature on the target [89]. A PDP can be constructed for categorical and continuous features as well as for classification and regression problems. However, the PDP assumes that the features are uncorrelated, which can be problematic for multidimensional prediction problems.

Permutation feature importance [101] estimates the importance of a particular feature to a trained model. It is the absolute difference in performance score when a real feature is replaced by a dummy feature; the more performance degrades, the more important that feature is to the model. The importance of the permutation depends on the model and the performance score chosen; any change in the performance score can change the ranking of the features. This method also cannot account for covariances between features.

Leave One Covariate Out (LOCO) [110] is a model-agnostic global and local feature importance method, similar to feature importance in Random Forests. However, unlike feature importance in Random Forests, the feature under consideration is not replaced by a dummy variable, but simply omitted. Both methods have in common that they ask *How good is the model without this feature?* The assumption behind this is that a feature is important for the model if the performance is significantly worse without this feature.

Maximum Mean Discrepancy (MMD-Critic) [106] distinguishes between representative samples of a class and outliers. Typical representative samples are called prototypes, and the outliers are called criticisms. The distinction between prototypes and outliers is intended to provide additional insight into the model.

Other methods for global model-agnostic diagnostics include accumulated local effects plots, H-statistics, and functional decomposition.

3.4.3.4 | Local Model-Agnostic Methods

Local model-agnostic methods explain individual predictions and are applicable to a wide range of machine learning models. Popular examples of methods in this category include Individual Conditional Expectation (ICE), Locally Interpretable Model-agnostic Explanation (LIME), anchors (scaled rules), SHapley Additive exPlanations (SHAP), and influence functions.

Individual conditional expectation (ICE) [103] is a refinement of PDP and accounts for the heterogeneity of individual data points. ICE disaggregates PDPs to illuminate individual conditional expectations from supervised models.

Local Interpretable Model-agnostic Explanation (LIME) [108] trains an interpretable model to approximate the predictions of the real model. LIME can locally explain text models from tree-based algorithms as well as computer vision models, such as deep neural networks. Later work has shown that random noise leads to instability in LIME-generated explanations [130; 131], leading to the development of LIME variants, including S-LIME [132] and DLIME [131].

Anchors (scoped rules) is another method developed by LIME authors [115]. In this method, IF-THEN rules are created to indicate which feature values anchor a prediction. Rules for rare classes or near the boundary of decision functions can become complex and sometimes ambiguous.

SHapley Additive ExPlanations (SHAP) is a model-agnostic method that allows for both global and local explanations and considers both structured and unstructured data [109]. SHAP indicates the contribution of a feature value to the difference between the actual prediction and the mean prediction. SHAP is based on Shapley values from game theory [133], dispersion activation features [104], and model intrinsic approaches from tree-based methods.

Influence functions [111] trace a prediction through the model and back to the training data to identify training points that are most responsible for a particular prediction. The influence function method can be applied to any model for which a second derivative exists. A derivative (synonym differentiation) exists, i. e., for neural networks. However, it is computationally intensive because the model must be re-trained when the training data changes.

Other local model-agnostic methods include individual conditional expectation curves (which can be used to generate partial dependence diagrams) and counterfactual explanations.

3.4.3.5 | Global Model-Specific Methods

Global model-specific methods describe the overall average behavior of a model for a given class of models. Methods in this category include Mean Decrease Impurity (MDI), Testing Concept Activation Vectors (TCAV), Soft Decision Trees, and TabNet. Mean Decrease Impurity (MDI) [102] explains the importance of features for tree ensembles. Testing Concept Activation Vectors (TCAV) is a global and local explanation method for computer vision models and tabular discrete data [114]. Soft decision trees use a decision tree to mimic the input-output function of a neural network [112]. In a soft decision tree, all leaf nodes contribute to the final decision with different probabilities [135]. For some leaf nodes, the soft decision tree allows a visual interpretation of

the neural network. However, not all learned filters are interpretable to the human eye. TabNet [295] uses sequential neural networks to mimic the logic of a decision tree on tabular data. Feature meanings provide global explanations, while heat maps provide local explanations. Instance-based feature selection can lead to confusion when local and global feature meanings contradict each other. Other global model-specific methods include Automatic Concept-Based Explanations (ACE) [117] and Deep Lattice Networks (DLN) [296]. Related to Decision Trees, but bringing in the aspect of symbolic AI, is the *Trepan Reloaded* method [297]. It uses ontologies to represent a network of information with logical relations and thus brings Domain Expert knowledge directly into the XAI system.

3.4.3.6 | Local Model-Specific Methods: Gradient-Based Explanations for Neural Networks

Local model-specific methods explain a particular prediction of a particular class of models. The most popular local model-specific methods are gradient-based neural network explanations, which we consider in this section as an example of this class.

Gradient-based neural network explanation methods use the gradient of a model to produce an explanation for a given input example and output class [123]. They are most applied to neural networks for image classification, for which they provide a visualization to highlight which regions of an input image were used to make a prediction.

Input-Level Gradient-Based Methods Gradient-based methods at the input level involve gradients or gradient-like calculations that lead from the output layer back to the input layer. So, the input layer refers to the level at input. Non-technical readers might want to know that the gradient is a derivative vector for a multivariate function, and the derivative of a function is the change of the function for a given input. Gradients are used, i. e., to fit neural networks to a dataset. The resulting explanation has the same number of pixels as the input image. Input layer approaches include saliency mapping, Guided Backpropagation, Deconvolutional Networks, SmoothGrad, Gradient \times Input, Layer-Wise Relevance Propagation, and DeepLIFT. All these approaches are computationally efficient but suffer from white noise caused by shattered gradients [298], which sometimes prevents the resulting explanations from appearing class-specific in practice. Saliency mapping is the original gradient-based explanation method for neural networks. Saliency mapping computes the gradient of the class score with respect to the input image [299]. DeconvNets [120] and Guided Backpropagation [137] are explanation methods developed independently that happen to be identical to saliency

mapping except for handling of the ReLU nonlinearities [128]. Saliency mapping passes explanation method sanity checks, while Guided Backpropagation does not, and may in fact function more like an edge detector than a model explanation [127]. SmoothGrad is another variant of saliency mapping that aims to reduce noise in explanations [113] but is not demonstrably more faithful to the model. The Gradient \times Input method is equivalent to Saliency Mapping, except that the saliency map is multiplied element by element with the input image to create the final visualization. It has been shown later that the Gradient \times Input method fails sanity checks [127]. Layer-Wise Relevance Propagation (LRP) [105] generates relevance values for the input pixels by iteratively distributing the final value across the layers of the neural network, starting with the output layer and working backwards to the input layer. Values greater than zero indicate that a particular pixel is relevant to the selected class. There are several variants of LRP. While LRP was not originally described as a gradient-based explanation method, it was later demonstrated [123] that ϵ -LRP is a variant of the gradient- $*$ input method, in which the gradient calculation is changed based on the ratio of output to input at each nonlinearity. Finally, Deep Learning Important Features (DeepLIFT) [104] provides explanations by estimating how much each neuron in a neural network is activated for an individual input compared to a reference input. The reference input is neutral (*foil*), while the individual input can be described as *fact* [300]. After the development of DeepLIFT, it was proved [123] that DeepLIFT computes backpropagation for a modified gradient function. Other input-level gradient-based methods include integrated gradients [107] and EXplanation Ranked Area Integrals (XRAI) [301]. For all these gradient-based methods one should keep in mind that the shattered gradients problem negatively affects the quality of the pixel importance values.

Output-Level Gradient-Based Methods In gradient-based explanations at the output layer, a gradient is computed that runs backwards from the output layer for only one or a few layers of the neural network without going all the way back to the input layer. Thus, the bare explanation has a smaller dimension than the input and must be upsampled before it is overlaid with the input to create the final explanation. Such an upsampling step is permissible because in a typical neural network the spatial relationship between output and input is preserved. Output-level approaches include Class Activation Mapping (CAM), Grad-CAM, and HiResCAM. Class Activation Mapping (CAM) is the fundamental method in this class [126]. CAM is based on a particular convolutional neural network architecture, where convolutional layers are followed by a global average pooling and a single fully connected layer, which provide the final predictions. A CAM explanation is obtained by multiplying the class-specific weights of the

final fully connected layer by the corresponding feature maps before the global average pooling step. CAM is a gradient-based method because these final weights represent the gradient of the class score with respect to the feature maps. The CAM method is trustworthy and guaranteed to highlight only regions the model used, but it has architecture restrictions. Gradient-weighted Class Activation Mapping (Grad-CAM) [24] aims to generalize CAM to other architectures. In Grad-CAM, the gradient of the class score is computed with respect to a given set of feature maps. Then, the gradient is averaged per feature and the averaged gradient is multiplied by the corresponding feature map. The aggregation of these weighted feature maps is the Grad-CAM explanation. The paper presenting Grad-CAM has been cited over 9,000 times, but unfortunately it has recently been shown that Grad-CAM is not faithful to the underlying model due to the gradient averaging step [291]. Grad-CAM's explanations highlight irrelevant regions of the input image that were not used for prediction, which can lead to misleading explanations that deviate significantly from the true behavior of the model [302]. HiResCAM [291] is a newer method that eliminates the inaccuracy of Grad-CAM. HiResCAM eliminates the gradient averaging step in Grad-CAM. By retaining the detailed gradient information and multiplying the gradients element by element with the corresponding feature maps, the relationship between the model explanation and the class evaluation is provably maintained, resulting in trustworthy class-specific explanations. The source code for HiResCAM is publicly available [here](#) and as part of [this package](#).

3.4.3.7 | Summary

Describing all the machine learning interpretability methods ever developed would require an entire textbook. Therefore, this section is not comprehensive, but rather is intended to provide representative examples of the major classes of interpretability methods. Table 3.17 gives a brief overview of common interpretability methods and summarizes the characteristics of each method.

3.4.4 | Materials and Methods

This section describes the search term, search results, inclusion and exclusion criteria, and research questions answered for each of the papers in this literature review.

3.4.4.1 | Literature Selection

Since we focus on medical data, we deliberately chose PubMed as our search database. Technical papers in this field, such as from the journal [Artificial Intelligence in Medicine](#),

Method / Taxonomy	Specific (S) or Agnostic (A)	Local (L) or Global (G)	Neural Networks	Computer Vision	Tabular Data	Year	No. Citations	Regr. (R) or Classif. (C)	Source Code Available
Partial Dependence Plots (PDP)[89]	A	G	No	No	Yes	2001	15545	R and C	Yes
Permutation Importance[101]	A	G	No	No	Yes	2010	15545	R and C	Yes
Mean Decrease Impurity[102]	S	G	No	No	Yes	2013	823	R and C	Yes
Individual Conditional Expectation[103]	A	L	Yes	No	Yes	2013	571	R and C	Yes
DeepLIFT (Deep Learning Important Features)[104]	S	L	Yes	Yes	No	2016	1629	C	Yes
Layer-Wise Relevance Propagation[105]	S	L	Yes	Yes	No	2016	2160	C	Yes
Maximum Mean Discrepancy - Critic[106]	A	G	Yes	Yes	No	2016	445	C	Yes
Gradient-weighted Class Activation Mapping[24]	S	L	Yes	Yes	No	2016	6758	C	Yes
Integrated Gradients[107]	S	L	Yes	Yes	No	2017	2017	C	Yes
Local Interpretable Model-agnostic Explanation (LIME)[108]	A	L	Yes	Yes	Yes	2017	5020	R and C	Yes
SHapely Additive exPlanations (SHAP)[109]	A	L and G	Yes	Yes	Yes	2017	5020	R and C	Yes
Leave One Covariate Out[110]	A	L	No	No	Yes	2017	274	R	Yes
Influence Functions[111]	A	L	Yes	Yes	No	2017	1377	C	Yes
Soft Decision Trees[112]	S	G	Yes	No	No	2017	357	C	Yes
SmoothGrad [113]	S	L	Yes	Yes	No	2017	867	C	Yes
Testing Concept Activation Vectors [114]	S	L and G	Yes	Yes	No	2018	583	C	Yes
Anchors [115]	A	L	Yes	Yes	Yes	2018	922	R and C	Yes
Representer Point Selection [116]	S	L	Yes	Yes	No	2018	105	C	Yes
Automatic Concept-based Explanations [117]	S	G	Yes	Yes	No	2019	157	C	Yes

Table 3.17: Overview of interpretability methods relevant to tabular and computer vision tasks, ordered by year of publication. Method relevance to neural networks, computer vision, and tabular data is indicated in the respective columns. The number of citations was derived from Google Scholar as of December 23, 2021. Links to the source code are provided via hyperlinks. We included methods that had more than 100 citations on Google Scholar, whose source code was publicly available, and that were optionally used in the review articles. An explanation of each method with advantages and limitations can be found in the supplementary material on [GitHub](#). Regr = Regression, Classif = Classification.

are also archived in PubMed, so we consider the database to be representative for our study purpose. We chose the search query (i. e., search term) that we applied to PubMed as follows: Within the title or abstract, we searched for terms associated with explainability, intelligibility or interpretability, combined with general terms that are related with machine learning and the medical domain. The search period covers the years from 2020 to 2022 and was conducted on March 7, 2022. We further applied the Preferred Reporting Items for Systematic Reviews [303] (PRISMA) guidelines. The aim of the literature search was to identify all papers from the last 20 years that met the following inclusion criteria:

- Used a supervised machine learning method,
- for a medical use case,
- with tabular or image data as input,
- and incorporating at least one explainability method.

Papers that focused on unsupervised learning (e.g., clustering), did not include a medical use case, or did not explicitly consider explainability were excluded. Borderline cases where it was not clear whether a paper should be included or not were discussed by the reviewers in separate meetings and concordantly accepted or rejected.

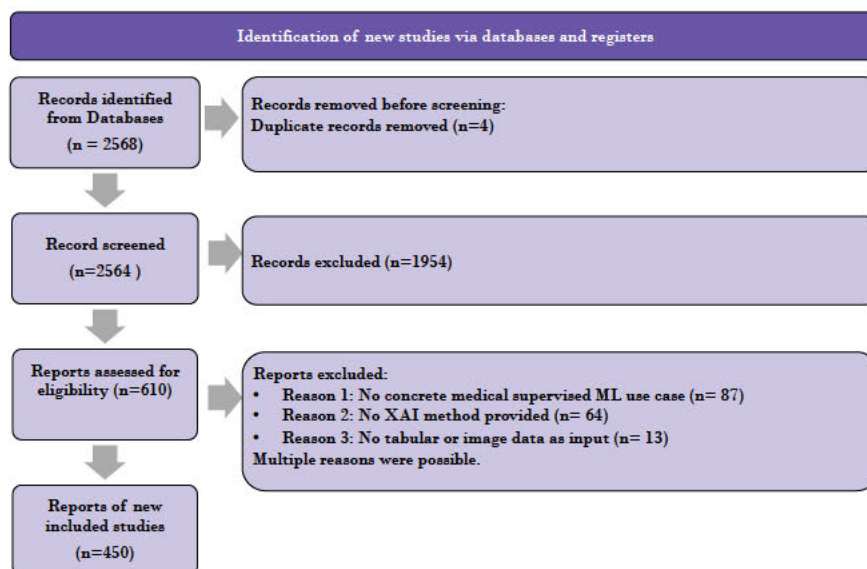


Figure 3.21: Flowchart of our PRISMA literature research in the PubMed database on 2022-03-07 based on the template of [303]. Out of 2568 papers, 450 have been finally included.

2,568 references were initially identified in the PubMed database on 2022-03-07. The search was limited to the title and abstract of the paper, meaning that machine learning and explainability had to be explicitly mentioned in these sections. After removing 4 duplicates, 2,564 references remained. Then, an author (J.A. or L.M.) applied the inclusion and exclusion criteria described above to each title and abstract, resulting in 610 references that were eligible for full-text screening. A common reason for exclusion at this stage was the lack of a true explainability method. For example, papers were excluded that used a black-box model but claimed that their model was explainable because a human could understand the dictionary definition of the input variable (e.g., "age"). We reviewed papers on time series data from electrocardiograms (ECGs) or electroencephalograms (EEGs), but unfortunately had to exclude all of these time series papers because none of them included an explainability method. Although our initial search term included papers from 2002 onward, the oldest paper that met the inclusion criteria was from 2008. The 610 references approved based on title and abstract were then subjected to full-text screening. At this stage, 160 references were excluded for at least one of the following reasons: no specific medical supervised ML use case (87), no XAI method (64), or no image or tabular data as input (13). There were 450 references left for data analysis with our 7 research questions. The Results section contains the analyses performed on these 450 references. The full PRISMA flowchart is shown in Figure 3.21.

3.4.4.2 | Literature Review

We evaluated each of the 450 final papers individually. For each paper, we determined which XAI methods were used and how they were described using 7 screening questions listed in Table 3.18. We did not consider model confidence estimation or model uncertainty as explanatory methods because they do not provide insight into how a model arrives at a prediction.

Data Synthesis We used Microsoft Forms to collect responses to our research questions and Python to aggregate the data according to our research questions. Each paper was reviewed by an author (J.A. or L.M.). We analyzed the Microsoft Forms data for all papers using Python 3.9⁴. All source code and raw data is available on the supplementary material on [GitHub](#).

The ML pipeline was rated 1, 2, or 3 according to the following criteria, which were agreed upon in advance:

- 1 = not described;

⁴<https://www.python.org/downloads/release/python-390/>

No.	Question	Answer options	Question type
1	Is there a concrete medical supervised ML use case?	Yes No	Single Choice
2	Which XAI method is used?	Several options including an open text field	Multiple Choice
3	From which data format is the input?	Tabular Image Audio Text	Single Choice
4	Who is potentially able to understand the XAI method?	Developers Medical Professionals Patients Other	Multiple Choice
5	How well is the ML pipeline described?	Not described Described Elaborately described	Single Choice
6	Is the source code provided?	Yes No Upon request	Single Choice
7	Is the data publicly available?	Yes No Upon request	Single Choice

Table 3.18: The research questions used were related to the 450 papers with a specific use case of supervised machine learning in medicine that used table or image data as input and applied at least one XAI method to explain ML predictions.

- 2 = described = the ML pipeline was mentioned and described briefly;
- 3 = elaborately described = the ML pipeline was described in detail, illustrated with a figure, or provided as publicly available code.

3.4.5 | Results

The results presented here are not always direct answers to the research questions but are primarily a combination of the information we obtained from the research questions. Therefore, a quick overview of the non-combined results is given in Figure 3.22

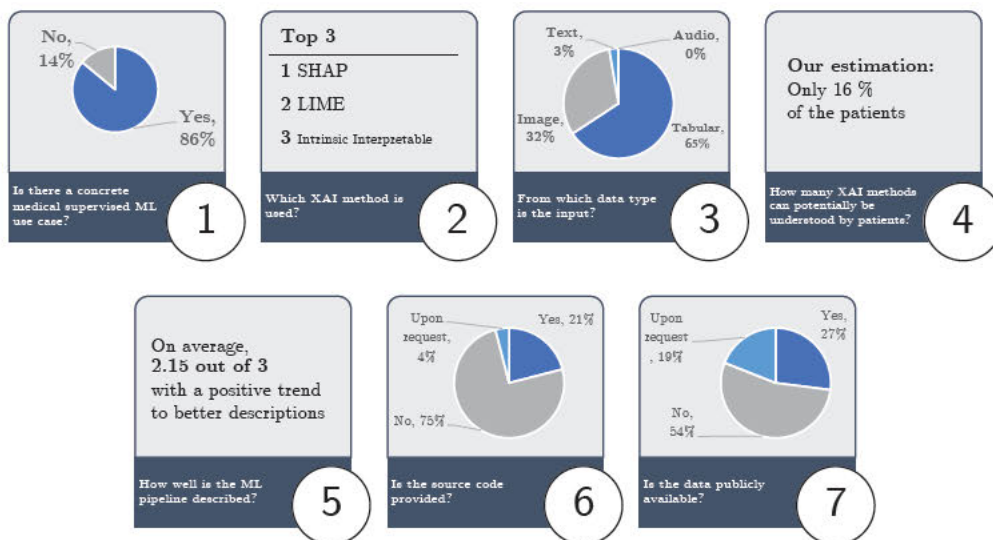


Figure 3.22: Brief overview of the results of the 7 research questions.

What is the publication rate of medical ML papers over time? The publication rate of medical ML papers increased over time, with 79 papers published between 2008 and

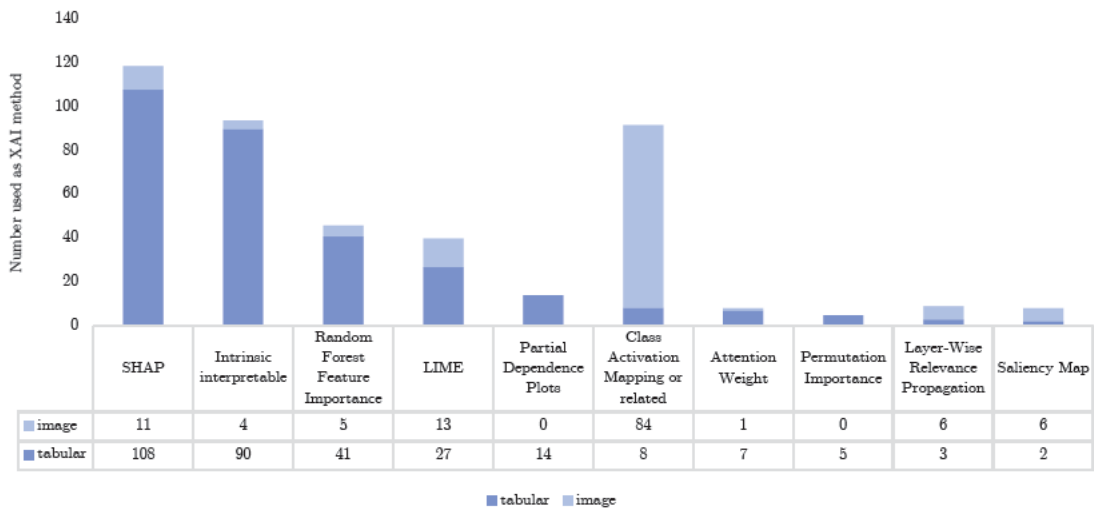


Figure 3.23: Number of XAI methods used for image and tabular data. At least one XAI method was considered for each paper. For tabular data, SHAP was by far the most popular XAI method; for image data, it was Grad-CAM. Methods that were used fewer than three times are not listed here.

2019, 108 more in 2020 only, and 200 in 2021. The 63 papers for 2022 were published within the first 67 days of the year. If we extrapolate this to 365 days, we expect about 343 publications in 2022.

What XAI methods are most used? Of the total 535 XAI methods used, 45 were developed by the authors of the use cases themselves and 490 were based on previously published methods such as SHAP, LIME or Grad-CAM. About 1.2 XAI methods were applied per paper. Figure 3.23 shows all XAI methods used in at least 3 papers, grouped by tabular data and image data. For tabular data, the most common XAI methods are SHAP, Random Forest Feature Importance, and intrinsic methods. For image data, the most used methods include class activation methods, SHAP, and LIME. Although we grouped gradient-based explanatory methods in the same category, we found that Grad-CAM was the most used method in this category.

Is tabular or image data more frequently used in medical ML papers? A total of 307 papers (68 %, (95 % CI [63.7 %, 72.5 %])) dealt with tabular data and the remaining 143 with image data. When grouping the type of input data by year, there is a clear trend towards greater use of image data over time. The ratio of tabular to image data was 20/80 in 2008-2019, while it increased to 36/64 in subsequent years. We think that this is likely

XAI Method	2008-2019	2020	2021	2022	Sum
SHAP	1	20	73	25	119
Intrinsic interpretable	33	25	34	2	94
Class Activation Mapping or related	7	23	40	22	92
Random Forest Feature Importance	9	13	19	5	46
LIME	5	6	24	5	40
Partial Dependence Plots	2		10	2	14
Layer-Wise Relevance Propagation	1	3	4	1	9
Attention Weight	2	3	2	1	8
Saliency Map	0	2	5	1	8
Permutation Importance	1	1	3	0	5
DeepLift	2	0	1	0	3
Sum (incl. all methods)	75	106	244	65	490

Figure 3.24: XAI methods used by year. Due to the small number of papers per year between 2008 and 2019, we have combined these years as one group. Note that for 2022, only papers published through 2022-03-07 were included.

due to the increasing availability of labeled image data and the increased computing power and availability of GPU clusters.

Who is potentially be able to understand the XAI method? On our opinion, patients are an important stakeholder group for machine learning applications in medicine and, on average, also have the greatest barriers to understanding due to their unfamiliarity with machine learning or medicine. To assess patient understanding of explainable ML models, we would ideally conduct a direct survey of patients. However, such a survey is beyond the scope of this work, so we instead made a subjective assessment of whether patients might be able to understand an XAI method. We considered a paper to be potentially understandable to patients if it met all the following criteria:

1. The output of the explainability method does not require a deeper medical understanding.
2. If variables are the output, they must be explained or self-explanatory.
3. If codes are the output (1=Female, 2=Male), they must be explained.
4. If color scales are the output, they must be explained with a legend and the meaning of the marginal values.

Only 16.4 % of the papers met all the above criteria to be considered potentially understandable by patients. A detailed overview of which of the explanatory methods presented in the papers could be understood by patients is given in Figure 3.25. Figure 3.26 again shows that explanation methods with image data as input were generally rated as better understood by patients, mainly due to the intuitive nature of heatmap displays of pixel relevance.

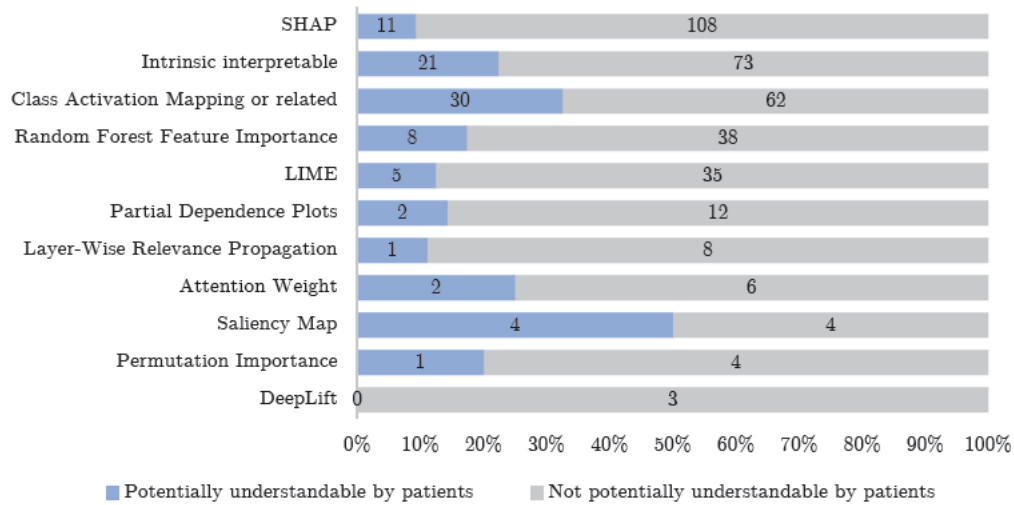


Figure 3.25: Which method of explanation is potentially understandable to patients? Each row sums to 100 %. The numbers within the bars indicate the number of papers that used this method in our review. For example, for the bar belonging to permutation meaning, we think that 1 in 5 papers (20 %) could be understood by patients.

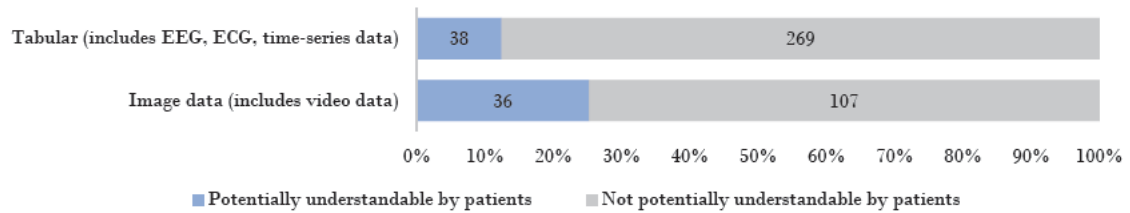


Figure 3.26: Which explanation method is potentially be understood by patients? Each row sums to 100 %. The numbers within the bars indicate the number of papers that fall into each category.

How well is the ML pipeline described? The mean ML pipeline score was 2.15 (0.60 std). From the aggregation of years in Figure 3.19, it appears that the description of the ML pipeline improves over time. This suggests that more weight was given to the ML pipeline in later work. Figure 3.27 shows the granularity of the ML pipeline description by input type.

Is the source code provided and is the data used publicly available? Overall, 75.6 % of all included papers do not make their source code available to reproduce the results. Surprisingly, the willingness to share source code was higher in 2020 (27.8 %) than in 2021 (15.5 %) or 2022 (20.6 %). We did not find any association between increasing years and the code sharing ratio. A chi-square test of independence showed that there was no

Year	# Papers	ML pipeline description	Code shared?			Data shared?		
			No	Upon request	Yes	No	Upon request	Yes
2008-2019	79	1.87	72,2%	5,1%	22,8%	58,2%	15,2%	26,6%
2020	108	2.15	68,5%	3,7%	27,8%	54,6%	18,5%	26,9%
2021	200	2.16	80,5%	4,0%	15,5%	52,0%	22,0%	26,0%
2022	63	2.30	76,2%	3,2%	20,6%	57,1%	12,7%	30,2%
Overall	450	2.15	73,6%	4,0%	20,4%	54,4%	18,7%	26,9%

Table 3.19: Assessment of ML pipeline and ratio of code and data sharing over time. For 2022, only publications through 2022-03-07 are considered. The column *ML pipeline description* refers to the research questions *How well is the ML pipeline described?* and is a mean score. The higher the score between 1 and 3, the more detailed the ML pipeline description.

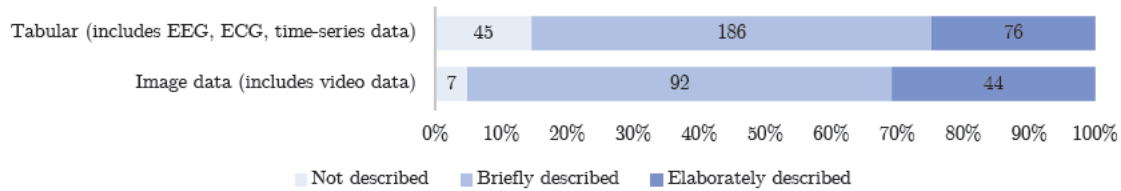


Figure 3.27: Granularity of machine learning description grouped by input type of data. We defined three categories: not described, briefly described, and elaborately described.

significant association between publication year and code sharing ratio, $\chi^2(6, N = 405) = 11.0, p = .09$. The willingness to share data is generally higher than the willingness to share code. In 2022, data availability was highest at 30.2%; in previous years, it was 27%. The increase in 2022 may also be since many papers have been published on the coronavirus radiograph use case and this data is publicly available [304]. For an overview, see Figure 3.28. Moreover, when a proprietary method was developed, willingness to share data was 4.4% points higher.

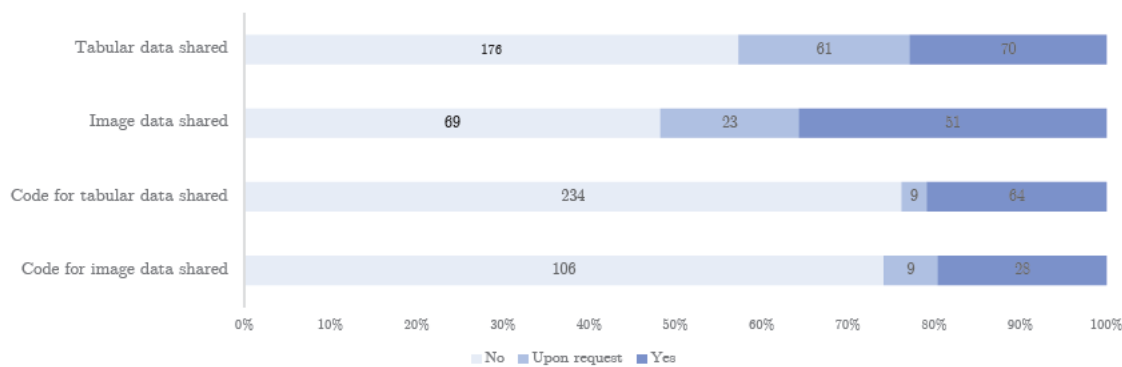


Figure 3.28: Availability of data and code to reproduce ML results. Each row sums to 100%. Note that the sum of the line *tabular data shared* is equal to *code for tabular data shared*. The same is true for the lines related to image data.

3.4.6 | Discussion

This paper has explored two main tasks. First, we have provided an overview of the taxonomy of machine learning explainability and described representative methods. The machine learning taxonomy has converged in recent years, although homonyms and synonyms still exist, and some technical terms are not used consistently. Second, we examined the last 20 years of medical machine learning publications in the PubMed database for their use cases, input types, AI comprehensibility, code and data sharing, and XAI methods used. The most popular methods are SHAP, LIME, and intrinsically interpretable methods. Most input data is structured, tabular data (65 %) or images (32 %). Text data is rare at 3 %, and we found no use case for our criteria for audio data. We estimate that 16 % of the explanatory methods reported in publications can be understood by patients. The description of machine learning pipelines has become more detailed over time, while data and code sharing has stagnated.

In this discussion section, we address the limitations of our study, provide recommendations on how to further improve the reproducibility and explainability of AI systems, address the interdisciplinary nature of medicine and machine learning, and mention challenges that may arise in the future and with the use of systems.

3.4.6.1 | Limitations of this review

We carefully discussed the search terms and the cases to be excluded before searching the databases to obtain as many precise hits as possible. However, because of the variable taxonomy in the ML community and the wide range of terms used in this inherently interdisciplinary field, we may not have identified all the papers that would have been relevant. We used the PubMed database because of our focus on medical use cases. PubMed is the largest medical database available. Google Scholar, i. e., has a more technical focus. In future work, we would like to extend our search on other databases. However, we do not expect our main findings to change. The final selection of 450 studies does not include papers on medical time-series data (e.g., ECG, EEG) or papers using support vector machines as a modeling framework because none of these papers considered explainability. Despite these limitations, we believe that our selection of relevant papers is large enough to derive representative conclusions. The results of our review are not free of subjectivity, especially when it comes to the evaluation of ML pipeline quality. Our assumptions regarding the comprehensibility of XAI methods to patients are also debatable. By setting criteria for when it is presumably understandable, a uniform assessment of the reviewed papers is ensured. However, the best estimator of

understandability is admittedly obtained from a representative survey with examples from the papers, which is out of scope as mentioned earlier.

3.4.6.2 | Recommendations to improve medical ML explainability and reproducibility

In reviewing the literature, we found that not all papers that include an explainability method explain it well. We have therefore developed several specific recommendations to improve the comprehensibility of medical XAI research.

First, the ML pipeline often requires more information about the split between training, validation, and testing. Some papers do not mention the data split at all, while others do not distinguish between validation and testing. We recommend at least mentioning the percentage split between training, validation, and test, and confirming that the final model's performance was calculated on the test set only. If cross-validation was used, we recommend indicating the robustness of the model by reporting the standard deviation of the performance metric across-validation folds. In some deep learning literature [305], a validation set is mostly used to avoid overfitting during training which is like maximizing robustness: Overfitting is indicated by a performance drop between training and validation fold, low robustness is indicated by a high variance between training and validation fold. We also recommend using the term "fold" when cross-validation was used, and "split" or "set" otherwise. A graphical concept for this can be seen in Figure 3.29. To give an example: there are 1,000 samples in a survey with 1,000 patients. We divide the samples into 800 for a training set and 200 for a test set. The test set is not used until the final evaluation. The training set is divided into 5 parts for 5-fold-cross-validation. 4 folds (=640 samples) are used for the first training while the 5th fold (160 samples) is tested. The average performance of the 5th folds is an estimator for the performance in the test set. The standard deviation in the test set, which results from the deviations between the test folds, is an estimator for the robustness of the model at deployment.

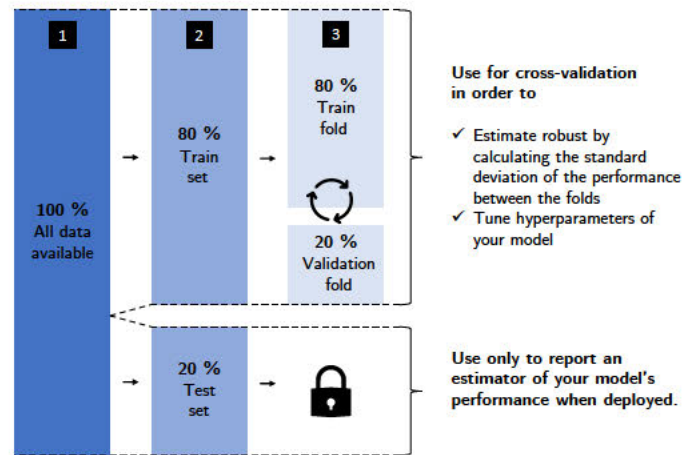


Figure 3.29: Our best practice recommendation for splitting all available data into training and validation folds and a test set. The 80-20 split is only a rule of thumb and can be adjusted depending on the amount of data available. The arrow circle refers to the shifting validation folds of the cross-validation. Note that we use the terms *validation* and *development* interchangeably. We denote *set* to data that is split into train and test sets. Within the train set, and when applying cross-validation, use the term *folds* to emphasize that this is not test, but validation data within the train set to estimate the model's generalization error. Av. = Available.

Second, "explaining the explanations" is often critical for XAI to fully deliver its intended benefits. For tabular data, it is important to explain what each variable means. Rather than using an abbreviation such as "BP," it is helpful to write out the full variable description: "blood pressure." Sometimes variables were not even abbreviated, but presented as numeric coding without a legend, e.g., "1", "2". When numeric coding is used, the meaning of each variable should be indicated, e.g., "1 = blood pressure," "2 = respiratory rate." For images, it is useful to indicate important anatomical structures or abnormalities with arrows so that non-radiologists can identify whether the model explanation overlaps with relevant parts of the medical image. It is also often helpful to recognize when an explanatory method is beyond a size that a human can easily understand. Decision Trees (DT) continue to be a popular tool in medical XAI. We have observed that Decision Trees are used both as an intrinsic method, i. e., the tree is the model and its own explanation, and in other work as a post-hoc method to approximate a more complicated model, such as a gradient boosting machine or a neural network. A problem arises when the decision tree becomes too large. We recommend avoiding Decision Trees with several dozen levels, as these are understandable to humans in theory but not in practice: The sum of all decision rules along a path might be too long to be comprehended in a clinical daily routine. However, in future, better explanation

methods must be also investigated. It is also important to clarify the splitting direction on the leaves. It is also relevant to consider the overall system of a use case with the three elements of the use case, the AI system, and the explanation of the AI system, rather than the XAI method separately. A suggestion from the community is to extend the descriptions of the XAI methods with standardized metadata that are uniform for all XAI methods and thus simplify the implementation and the technical access, analogous to FAIR [306] (Findable, Accessible, Interoperable, Reusable) principle [307].

Finally, to facilitate replication and faster progress of medical ML research, we encourage increased de-identification (anonymization of personal data) and sharing of datasets, as it is often difficult to build directly on medical ML research when a new dataset needs to be created from scratch.

3.4.6.3 | Understandability of an XAI method in medical ML is related to medical knowledge

In medical ML applications, medical knowledge is often useful to understand the results of an XAI method. Using image-based heatmap XAI methods as an example, we can consider different degrees of understanding. A general reader can reach a basic level of understanding, which we define as awareness that the highlighted pixels are relevant to the prediction. A physician who is not a radiologist would be able to reach an intermediate level of understanding, meaning that he or she is able to recognize organs and major abnormalities in the underlying medical image and consider how these relate to the relevance of the pixels. A radiologist would eventually be able to recognize even subtle anomalies in the medical image and assess their relationship to XAI pixel relevance. This means on the one hand that the degree of comprehensibility is essentially a subjective assessment, and on the other hand that the potential of comprehensibility depends on the background knowledge of the viewer. The more specific the AI application, the higher the dependency on domain knowledge for the comprehensibility of the XAI system.

3.4.6.4 | Challenges

Some enthusiasts believe that the use of black-box ML systems is unproblematic [308], while the most conservative work argues that not even existing explainable ML methods are sufficiently understandable to justify the use of ML in a clinical setting, since explainable ML cannot confirm the correctness of a decision [309].

We take an intermediate perspective in which we believe that explainable ML has the potential to improve clinical care in certain circumstances. In our opinion, any deployment of a medical ML model should involve close collaboration between medical professionals,

ML engineers, software developers, and computer security experts. Medical professionals have the deepest understanding of the model's explanations and can confirm whether a model's behavior appears medically appropriate. Only explanatory methods that are demonstrably faithful to the model should be used. Bias and fairness metrics should be calculated to ensure that the models used do not exhibit discriminatory behavior. Further, we think that the model must be protected from unauthorized access, and ML experts must be available to update the model in the event of a concept or data mismatch. In many countries, regulatory approvals are required for newly trained models.

Deploying a model is no guarantee that it will be used clinically. We think that the likelihood that a model will impact clinical care is greatest when the program's user interface for using the model has been carefully developed with significant input from medical professionals and when the model's outputs can be seamlessly integrated into existing software tools and workflows. Explainable ML methods with demonstrable guaranteed fidelity to the underlying model have the potential to improve the quality of medical ML models and prevent the use of possibly critical, biased, or ineffective models. The more medical ML research incorporates explainable techniques, the more clinical relevance it could achieve.

Supplementary Information

Supplementary materials, such as detailed descriptions of the XAI methods, as well as the Python code to replicate numbers, figures and tables are available on github.com/joa24jm/literature_review.

Additional Information

The authors declare no competing interests.

3.5 | 7 observational mHealth studies and 10 years of experience: Can ignoring groups in Machine Learning pipelines lead to overestimation of model performance? Analyses of group-wise validation as well as baseline and concept-drift considerations.

- **Authors** | Allgaier, Johannes and Pryss, Rüdiger
- **Under Review in** | Nature Communications Medicine

Abstract

Background. Machine learning (ML) models are evaluated in a test set to estimate model performance after deployment. The design of the test set is therefore of importance because if the data distribution after deployment differs too much, the model performance decreases. At the same time, the data often contains undetected groups. For example, multiple assessments from one user may constitute a group, which is usually the case in mHealth scenarios.

Methods. In this work, we evaluate a model's performance using several cross-validation train-test-split approaches, in some cases deliberately ignoring the groups. By sorting the groups (in our case: users) by time, we additionally simulate a concept drift scenario for better external validity. For this evaluation, we use 7 longitudinal mHealth datasets, all containing Ecological Momentary Assessments (EMA). Further, we compared the model performance with baseline heuristics, questioning the essential utility of a complex ML model.

Results. Hidden groups in the dataset leads to overestimation of ML performance after deployment. For prediction, a user's last completed questionnaire is a reasonable heuristic for the next response, and potentially outperforms a complex ML model. Because we included 7 studies, low variance appears to be a more fundamental phenomenon of mHealth datasets.

Conclusion. The way mHealth-based data are generated by EMA leads to questions of user and assessment level and appropriate validation of ML models. Our analysis shows that further research needs to follow to obtain robust ML models. In addition, simple heuristics can be considered as an alternative for ML. Domain experts should be consulted to find potentially hidden groups in the data.

3.5.1 | Introduction

When machine learning models are applied to medical data, an important question is whether the model learns subject-specific characteristics (not desired effect) or disease-related characteristics (desired effect) between an input and output. A recent paper by Kunjan et al. [310] describes this very well at the example of classification and EEG disease diagnosis. In the Kunjan paper, this is discussed using different variants of cross-validation. It is well shown that the type of validation can cause extreme differences. Older work has evaluated different cross-validation techniques on datasets with different recommendations for the number of optimal folds [86; 311]. We transfer and adapt this idea to mHealth data and the application of machine-learning-based classification and raise new questions about this. To this end, we will briefly explain the background. Using simple, understandable models rather than complex black box models is a clamor of Rudin et. al., which motivates us to evaluate simple heuristics against complex models [55]. The Cross-Industry Standard Process for Data Mining (CRISP-DM) highlights the importance of subject matter experts to get familiar with a dataset [312]. In turn, familiarity with the dataset is necessary to detect hidden groups in the dataset. In our mHealth use cases, one app user that fills out more several questionnaires constitutes a group.

We have developed numerous applications in mobile health in recent years (e.g., [8; 313]) and the issue of disease-related or subject-specific characteristics is particularly pronounced in these applications. mHealth applications very often use the principles of Patient-reported Outcome Measures (PROMs) or/and Ecological Momentary Assessments (EMAs) [30]. EMAs have the major goal that users record symptoms several times a day over a longer period. As a result, users of an mHealth solution generate longitudinal data with many assessments. Since not all users respond equally frequently in the applications (as shown by many applications that have been in operation for a long time [314]), the result is a very different number of assessments per user. Therefore, the question arises in the application of machine learning, how the actual learning takes place. In learning, should we group the ratings per user so that a user only appears in either the training set or the testing set, which is correct by design. Or, can we accept that a user's ratings appear in both the training and test sets, since users with many ratings have such a high variance in ratings. Finally, individual users may undergo concept drift in the way they answer questions in many assessments over a long period of time. In such a case, the question also arises as to whether it makes sense to use an individual's ratings separately in the training and testing sets.

In this context, we also see another question as relevant that is not given enough attention: What is an appropriate baseline for a machine learning outcome in studies? As mentioned earlier, some mHealth users fill out thousands of assessments, and do so for years. In this case, there may be questions about whether a previous assessment can reliably predict the next one, and the use of machine learning may be poorly targeted.

With respect to the above research questions, we use another component to further promote the results. We selected seven studies from the pool of developed apps that we will use for the analysis of this paper. Since a total of 7 studies are used, a more representative picture should emerge. However, since the studies do not all have the same research goals, classification tasks need to be found per app to make the overall results comparable. The studies also do not all have the same duration. Even though the studies are not always directly comparable, the setting is very promising as the results will show in the end. Before deriving specific research questions against this background, related work and technical background information will be briefly discussed.

3.5.1.1 | Existing train-test-split approaches

Within cross-validation, there exist several approaches on how to split up the data into folds and validate them, such as the k -fold approach with k as the number of folds in the training set. Here, $k - 1$ folds form the training folds and one fold is the validation fold [184]. One can then calculate k performance scores and their standard deviation to get an estimator for the performance of the model in the test set, which itself is an estimator for the model's performance after deployment (see also Fig. 3.31). In addition, there exist the following strategies:

- (Repeated) stratified k -fold, in which the target distribution is retained in each fold, which can also be seen in Figure 3.30. After shuffling the samples, the stratified split can be repeated [311].
- Leave-*one*-out cross-validation [315], in which the validation fold contains only *one* sample while the model has been trained on all other samples.
- Leave- p -out cross-validation, in which $\binom{n}{p}$ train-test-pairs are created with n equals number of assessments (synonym *sample*) [316].

These approaches, however, do not always focus on samples that might belong to our mHealth data peculiarities. To be more specific, they do not account for users (syn. groups, subjects) that generate daily assessments (syn. samples) with a high variance.

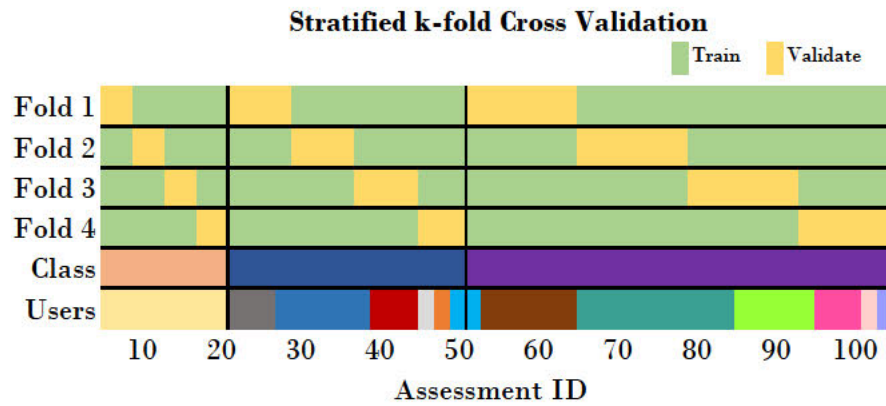


Figure 3.30: Illustration of train-validate split for stratified 4-fold cross-validation. While this approach retains the class distribution in each fold, it still ignores user groups. Each color represents a different class or user id.

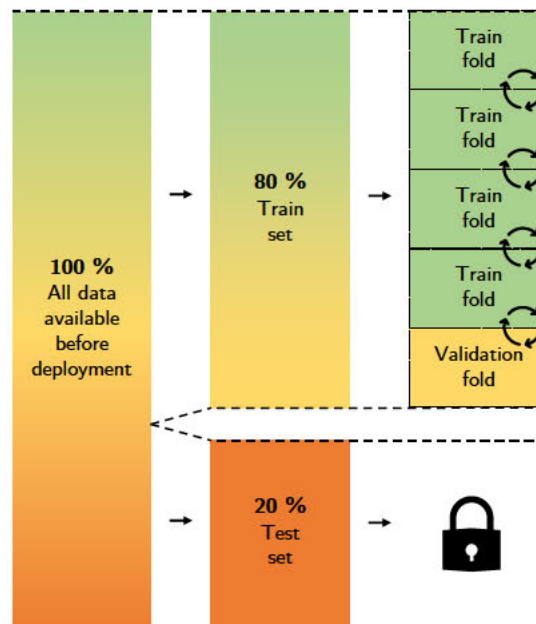


Figure 3.31: Schematic visualisation of the steps required to perform a k -fold cross-validation, here with $k = 5$.

3.5.1.2 | Related Work

Cawley et. al. also address the question of how to minimize the error in the estimator of performance in ground truth. Using synthetic data sets, they argue that overfitting a model is as problematic as selection bias in the training data [317]. However, they do not address the phenomenon of groups in the data. Refaeilzadeh et. al. give an overview of common cross-validation techniques such as leave-one-out, repeated k-fold, or hold-out validation [318]. They discuss pros and cons of each kind and mention an *underestimated performance variance* for repeated k-fold cross-validation, but they also do not address the problem with (unknown) groups in the dataset [318]. Schratz et. al. focus on spatial auto correlation and spatial cross-validation rather than on groups and splitting approaches [319]. Spatial cross-validation is sometimes also referred to as block cross-validation [320]. They observe large performance differences in the use or non-use of spatial cross-validation. By random sampling of train and test samples, a train and test sample might be too close to each other on a geographical space, which induces a selection bias and thus an overoptimistic estimate of the generalization error. They then use spatial cross-validation. We would like to briefly differentiate between *space* and *group*. Two samples belong to the same space if they are geographically close to each other. They belong to the same group if a domain expert assigns them to a group. In our work, multiple assessments belonging to one user form a group. Meyer et. al. also evaluate using a spatial cross-validation approach, but also add a time dimension using Leave-Time-Out cross-validation where samples belong to one fold if they fall into a specific time range [321]. This leave-time-out approach is like our *time-cut* approach, which will be introduced in the methods section. Yet, we are not aware of any related approach on mHealth data like the one we are pursuing in this work.

3.5.1.3 | Research questions

As written at the beginning of the introduction, we want to evaluate how much the model's performance depends on specific users (syn. *subjects, patients, persons*) that are represented several times within our dataset, but with a varying number of assessments per user. From previous work, we already know that so-called power-users with many more assessments than most of the other users have a high impact on the models training procedure [31]. We would further like to investigate whether a simple heuristic can outperform complex ensemble methods. Simple heuristics are interesting because they are easy to understand, have a low maintenance requirement, and have low variance, but also generate high bias.

Technically, across studies (i.e., across the seven studies), we investigate simple heuris-

tics at the user and assessment level and compare them to tree-based non-tuned ML ensembles. Tree-based methods have already been proven in the literature on the specific mHealth data used, that is why we use only tree-based methods. The reason for not tuning these models is that we want to be more comparable across the used studies. With these levels of consideration, we would like to elaborate on the following research questions:

- RQ1: What is the variance in performance when using different splitting methods for train and test set of mHealth data?
- RQ2: In which cases is the development, deployment and maintenance of a ML model compared to a simple baseline heuristic worthwhile when being used on mHealth data?

3.5.2 | Materials and Methods

In this section, we first describe how Ecological Momentary Assessments work and how they differentiate from assessments that are collected within a clinical environment. Second, we present the studies and ML use cases for each dataset. Next, we introduce the non-ML baseline heuristics and explain the ML preprocessing steps. Finally, we describe the splitting approaches at the user- and assessment levels.

3.5.2.1 | Ecological Momentary Assessments

Within this context, *ecological* means "within the subject's natural environment", and *momentary* "within this moment" and ideally, in real time [2]. Assessments collected in research or clinical environments may cause recall bias of the subject's answers and are not primarily designed to track changes in mood or behavior longitudinally. Ecological Momentary Assessments (EMA) thus increase validity and decrease recall bias. They are suitable for asking users in their daily environment about their state of being, which can change over time, by random or interval time sampling. Combining EMAs and mobile crowdsensing sensor measurements allows for multi-modal analyses, which can gain new insights in, e.g., chronic diseases [30; 31]. The datasets used within this work have EMA in common and are described in the following subsection.

3.5.2.2 | The ML use cases

From ongoing projects of our team, we are constantly collecting mHealth data as well as Ecological Momentary Assessments [4; 8; 322; 323]. To investigate how the machine

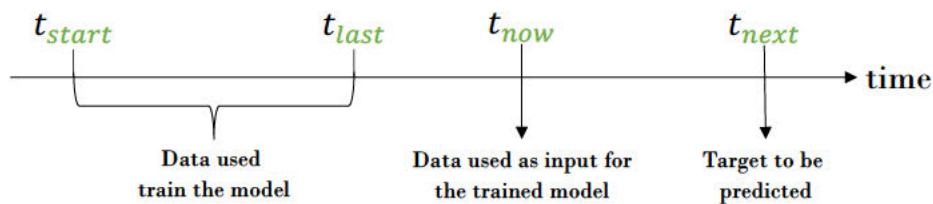


Figure 3.32: Schematic representation of the relevant four points in time for the understanding of the pipeline. At time t_{start} , the first assessment is given; t_{last} is the last known assessment used for training, whereas t_{now} is the currently available assessment as input for the classifier and the target is predicted at time t_{next} .

learning performance varies based on the splits, we wanted different datasets with different use cases. However, to increase comparability between the use cases, we created multi-class classification tasks.

We train each model using historical assessments, the oldest assessment was collected at time t_{start} , the latest historical assessment at time t_{last} . A current assessment is created and collected at time t_{now} , a future assessment at time t_{next} . Depending on the study design, the actual point of time t_{next} may be in some hours or in a few weeks from t_{now} . For each dataset and for each user, we want to predict a feature (synonym, a question of an assessment) at time t_{next} using the features at time t_{now} . This feature at time t_{next} is then called the target. For each use case, a model is trained using data between t_{start} and t_{last} , and given the input data from t_{now} , it predicts the target at t_{next} . Figure 3.32 gives a schematic representation of the relevant points of time t_{start} , t_{last} , t_{now} , and t_{next} . To increase comparability between the approaches, we used the same model architecture with the same pseudo-random initialisation. The model is a Random Forest classifier with 100 trees and the Gini impurity as the splitting criterion. The whole coding was in Python 3.9, using mostly *scikit-learn*, *pandas* and [Jupyter Notebooks](#). Details can be found on [GitHub](#) in the supplementary material.

The included apps and studies in more detail The following section provides an overview of the studies, the available datasets with characteristics, and then describes each use case in more detail. A brief overview is given in Table 3.20 with baseline statistics for each dataset in Table 3.21.

To provide some more background info about the studies: The analyses happen with all apps on the so-called **EMA questionnaires** (synonym: assessment), i.e., the questionnaires that are filled out multiple times in all apps and the respective studies. This can

3.5. 7 observational mHealth studies and 10 years of experience: Can ignoring groups in Machine Learning pipelines lead to overestimation of model performance? Analyses of group-wise Chapter 3. Results validation as well as baseline and concept-drift considerations.





Mobile Application	 Track Your Tinnitus TYT	 Corona Check CC	 Corona Health CH	 Unification of Treatments and Interventions for Tinnitus Patients UNITI
Studies Involved & Background	TYT was launched in 2014 to find out more about daily tinnitus fluctuations and since then has been a longitudinal observational study that we initiated without a project background using an iOS and Android app in the official app stores.	During the Covid 19 pandemic, Covid testing, and knowledge of the virus were scarce. Thus, we developed the CC app to provide feedback based on their reported symptoms.	<ul style="list-style-type: none"> • Mental health for <ul style="list-style-type: none"> • adults (CHA) • adolescents (CHY) • Physical health for <ul style="list-style-type: none"> • adults (CHP) • Stress (CHS) This app contains mental and physical health studies where user behavior can be tracked during the Covid pandemic.	UNITI wants to develop a model that enables patient-specific treatment of tinnitus.
Project Partners	Tinnitus Research Initiative	Bavarian State Office for Health and Food Safety	Robert Koch Institute	European Union's Horizon 2020 Research and Innovation Programme

Table 3.20: Overview of the mobile applications and the studies involved in this project: TrackYourTinnitus [4], Corona Check [324], Corona Health [8], and Unification of Treatments and Interventions for Tinnitus Patients [323].

Dataset	No. of users	No. of assessments	First assesment from	Dataset span	Ø Age (Std)	Ratio m/f/d	% rate of GER users
TYT	3303	110983	2013-07-18	9,20	45.0 (14.4)	67/33/00	n. A.
CC	13763	89659	2020-04-08	2,48	32.7 (18.0)	59/39/01	36
CHA	1474	11081	2020-07-21	2,19	41.2 (13.9)	54/45/01	98
CHP	953	5661	2020-07-28	2,17	41.8 (15.2)	63/37/00	98
CHY	111	630	2020-08-08	2,14	15.2 (1.6)	51/47/01	n. A.
CHS	374	3845	2020-12-19	1,78	40.7 (13.9)	65/34/01	98
UNITI	763	32443	2021-04-13	1,46	53.0 (12.7)	57/43/00	54

Table 3.21: Baseline statistics and overview of the datasets used. Ratio m/f/d is the sex ratio of male, female and diverse users. The dataset span is given in years. GER = German.

happen several times a day (e.g., for the tinnitus study TrackYourTinnitus (TYT)) or at weekly intervals (e.g., studies in the Corona Health (CH) app). Nevertheless, the analysis happens on the recurring questionnaires, which collect symptoms over time and in the real environment through unforeseen (i.e., random) notifications.

The TrackYourTinnitus (TYT) dataset has the most filled out assessments with more than 110,000 questionnaires as by 2022-10-24. The Corona Check (CC) study has the most users. This is because each time an assessment is filled out, a new user can optionally be created. Notably, this app has the largest ratio of non-German users and the youngest user group with the largest standard deviation. The Corona Health (CH) app with its studies *Mental health for adults, adolescents and physical health for adults* has the highest proportion of German users because it was developed in collaboration with the Robert Koch Institute and was primarily promoted in Germany. Unification of treatments and Interventions for Tinnitus patients (UNITI) is a European Union wide project, which overall aim is to deliver a predictive computational model based on existing and longitudinal data [323].

TrackYourTinnitus (TYT) With this app, it is possible to record the individual fluctuations in tinnitus perception. With the help of a mobile device, users can systematically measure the fluctuations of their tinnitus. Via the [TYT website](#) or the app, users can also view the progress of their own data and, if necessary, discuss it with their physician. The ML task at hand is a classification task with target variable *Tinnitus distress* at time t_{now} and the questions from the daily questionnaire as the features of the problem. The target's values range in $[0, 1]$ on a continuous scale. To make it a classification task, we created bins with step size of 0.2 resulting in 5 classes. The features are *perception, loudness, and stressfulness* of tinnitus, as well as the current *mood, arousal and stress level* of a user, the *concentration level* while filling out the questionnaire, and *perception of the worst tinnitus symptom*. A detailed description of the features was already done in previous works [52]. Of note, the time delta of two assessments of one user at t_{next} and t_{now} varies between users. Its median value is 11 hours.

Unification of Treatments and Interventions for Tinnitus Patients (UNITI) The overall goal of UNITI is to treat the heterogeneity of tinnitus patients on an individual basis. This requires understanding more about the patient-specific symptoms that are captured by EMA in real time.

The use case we created at UNITI is like that of TYT. The target variable *encumbrance*, coded as *cumbersness*, which was also continuously recorded, was divided into an ordinal scale from 0 to 1 in 5 steps. Features also include momentary assessments of the user during completion, such as *jawbone, loudness, movement, stress, emotion*, and questions

about momentary tinnitus. The data was collected using our mobile apps [313]. Here, of note: on average, the median time gap between two assessment is 24 hours for each user.

Corona Check (CC) At the beginning of the COVID-19 pandemic, it was not easy to get initial feedback about an infection, given the lack of knowledge about the novel virus and the absence of widely available tests. To assist all citizens in this regard, we launched the mobile health app Corona Check together with the *Bavarian State Office for Health and Food Safety* [324].

The Corona Check dataset predicts whether a user has a Covid infection based on a list of given symptoms [38]. It was developed in the early pandemic back in 2020 and helped people to get quick estimate for an infection without having an antigen test. The target variable has four classes:

- Suspected coronavirus (COVID-19) case
- Symptoms, but no known contact with confirmed corona case
- Contact with confirmed corona case, but currently no symptoms
- Neither symptoms nor contact

The features are a list of Boolean variables, which were known at this time to be typically related with a Covid infection, such as fever, a sore throat, a runny nose, cough, loss of smell, loss of taste, shortness of breath, headache, muscle pain, diarrhea, and general weakness. Depending on the answers given by a user, the application programming interface returned one of the classes. The median time gap of two assessments for the same user is 8 hours on average with a much larger standard deviation of 24.6 days.

Corona Health | Mental health for adults (CHA) The last four use cases are all derived from a bigger COVID-19 related mHealth project called *Corona Health* [8; 325]. The app was developed in collaboration with the Robert Koch-Institute and was primarily promoted in Germany, it includes several studies about the mental or physical health, or the stress level of a user. A user can download the app and then sign up for a study. He or she will then receive a baseline one-time questionnaire, followed by recurring follow-ups with between-study varying time gaps. The follow-up assessment of CHA has a total of 159 questions including a full PHQ9 questionnaire [326]. We then used the nine questions of PHQ9 as features at t_{now} to predict the level of depression for this user for t_{next} . Depression levels are ordinally scaled from *None* to *Severe* in a total of 5 classes. The median time gap of two assessments for the same user is 7.5 days. That is, the models predict the future in this time interval.

Corona Health | Mental health for adolescents (CHY) Similar to the adult cohort, the mental health of adolescents during the pandemic and its lock-downs is also captured by our app using EMA.

A lightweight version of the mental health questionnaire for adults was also offered to adolescents. However, this did not include a full PHQ9 questionnaire, so we created a different use case. The target variable to be classified on a 4-level ordinal scale is *perceived dejection* coming from the PHQ instruments, features are a subset of quality of live assessments and PHQ questions, such as concernment, tremor, comfort, leisure quality, lethargy, prostration, and irregular sleep. For this study, the median time gap of two follow up assessments is 7.3 days.

Corona Health | Physical health for adults (CHP) Analogous to the mental health of adults, this study aims to track how the physical health of adults changes during the pandemic period.

Adults had the option to sign up for a study with recurring assessments asking for their physical health. The target variable to be classified asks about the constraints in everyday life that arise due to physical pain at t_{next} . The features for this use case include aspects like sport, nutrition, and pain at t_{now} . The median time gap of two assessments for the same user is 14.0 days.

Corona Health | Stress (CHS) This additional study within the Corona Health app asks users about their stress level on a weekly basis. Both features and target are assessed on a five-level ordinal scale from *never* to *very often*. The target asks for the ability of stress management, features include the first nine questions of the perceived stress scale instrument [327]. The median time gap of two assessments for the same user on average is 7.0 days.

3.5.2.3 | Baseline heuristics instead of complex ML models?

We also want to compare the ML approaches with a baseline heuristic (*synonym: Baseline model*). A baseline heuristic can be a simple ML model like a linear regression or a small Decision Tree, or alternatively, depending on the use case, it could also be a simple statement like "**The next value equals the last one**". The typical approach for improving ML models is to estimate the generalization error of the model on a benchmark data set when compared to a baseline heuristic. However, it is often not clear, which baseline heuristic to consider, i.e.: The same model architecture as the benchmark model, but without tuned hyperparameters? A simple, intrinsically explainable model with or

without hyperparameter tuning? A random guess? A naive guess, in which the majority class is predicted? Since we have approaches on a user-level (i.e., we consider users when splitting) and on an assessment-level (i.e., we ignore users when splitting), we also should create baseline heuristics on both levels. We additionally account for within-user variance in Ecological Momentary Assessments by averaging a user's previously known assessments. *Previously known* here means that we calculate the mode or median of all assessments of a user that are older than the given timestamp. In total, this leads to four baseline heuristics (user-level latest, user-level average, assessment-level latest, assessment-level average) that do not use any machine learning but simple heuristics. On the assessment-level, the latest known target or the mean of all known targets so far is taken to predict the next target, no matter of the user-id of this assessment. On the user-level, either the last known, or median, or mode value of *this user* is taken to predict the target. This, in turn, leads to a cold-start problem for users that appear for the first time in a dataset. In this case, either the last known, or mode, or median of all assessments that are known so far are taken to predict the target.

3.5.2.4 | ML Preprocessing

Before the data and approaches could be compared, it was necessary to homogenize them. In order for all approaches to work on all data sets, at least the following information is necessary: `Assessment_id`, `user_id`, `timestamp`, `features`, and the target. Any other information such as GPS data, or additional answers to questions of the assessment, we did not include into the ML pipeline. Additionally, targets that were collected on a continuous scale, had to be binned into an ordinal scale of five classes. For an easier interpretation and readability of the outputs, we also created label encodings for each target. To ensure consistency of the pre-processing, we created helper utilities within Python to ensure that the same function was applied on each dataset. For missing values, we created a user-wise missing value treatment. More precisely, if a user skipped a question in an assessment, we filled the missing value with the mean or mode (*mode* = most common value) of all other answers of this user for this assessment. If a user had only one assessment, we filled it with the overall mean for this question.

For each dataset and for each script, we set random states and seeds to enhance reproducibility. For the outer validation set, we assigned the first 80 % of all users that signed up for a study to the train set, the latest 20 % to the test set. To ensure comparability, the test users were the same for all approaches. We did not shuffle the users to simulate a deployment scenario where new users join the study. This would also add potential concept drift from the train to the test set and thus improve the simulation quality.

For the cross-validation within the training set, which we call internal validation, we chose a total of 5 folds with 1 validation fold. We then applied the four baseline heuristics (on user level and assessment level with either latest target or average target as prediction) to calculate the within-train-set performance standard deviation and the mean of the weighted F1 scores for each train fold. The mean and standard deviation of the weighted F1 score are then the estimator of the performance of our model in the test set.

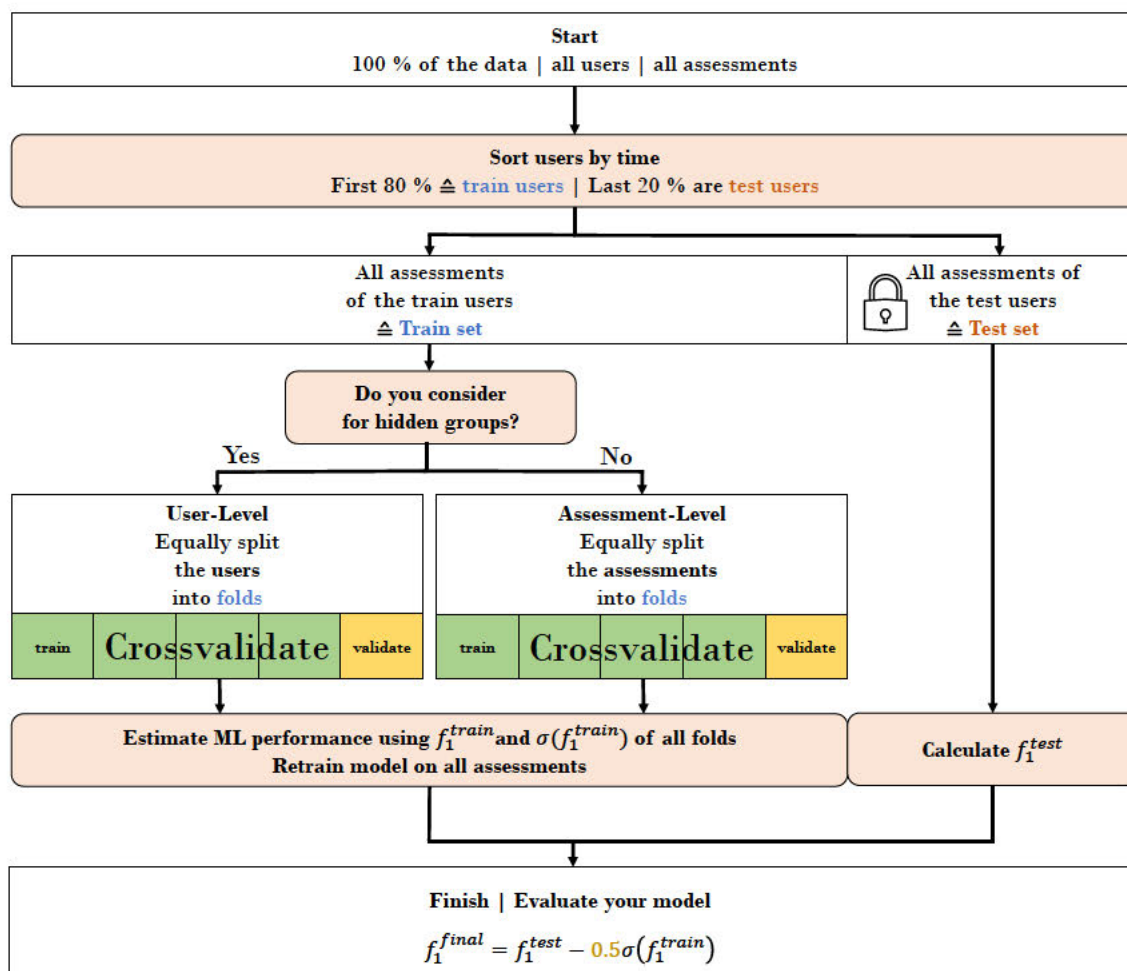


Figure 3.33: Adapted cross-validation schema with a user-based and assessment-based approaches. f_1^{train} conforms the average f_1 score of validation folds. f_1^{test} equals the f_1 score in the test set. $\sigma(f_1^{train})$ conforms the standard deviation of the f_1 scores of the validation folds and can be interpreted as an estimator of the generalization error of the model in the test set. The final score of a model is therefore smaller, the larger the standard deviation of the validation folds in the training set.

We call one approach superior to another if the final score is higher. The final score to evaluate an approach is calculated as:

$$f_1^{final} = f_1^{test} - 0.5\sigma(f_1^{train})$$

If the standard deviation between the folds during training is large, the final score is lower. The test set must not contain any selection bias against the underlying population. The pre-factor of the standard deviation σ with 0.5 has been chosen arbitrarily. It should be set higher the more important the generalization error of the model is, i.e., models with high performance variance between validation folds during training will receive an even lower final score.

3.5.2.5 | Splitting approaches related to EMA

To precisely explain the splitting approaches, we would like to differentiate between the terms *folds* and *sets*. We call a chunk of samples (synonym: assessments, filled out questionnaires) a *set* on the outer split of the data, for which we cut-off the final test *set*. However, within the training set, we then split further to create training and validation *folds*. That is, using the term *fold*, we are in the context of cross-validation. When we use the term *set*, then we are in the outer split of the ML pipeline. Figure 3.33 visualizes this approach. Following this, we define 4 different approaches to split the data. For one of them we ignore the fact that there are users, for the other three we do not. We call these approaches *user-cut*, *average-user*, *user-wise* and *time-cut*. All approaches have in common that the first 80 % of all users are always in the training set and the remaining 20 % are in the test set. A schematic visualization of the splitting approaches is shown in Fig. 3.34. Within the training set, we then split on user-level for the approaches *user-cut*, *average-user* and *user-wise*, and on assessment-level for the approach *time-cut*.

In the following section, we will explain the splitting approaches in more detail. The *time-cut* approach ignores the fact of given groups in the dataset and simply creates validation folds based on the time the assessments arrive in the database. In this example, the month, in which a sample was collected, is known. More precisely, all samples from January until April are in the training set while May is in the test set. The *user-cut* approach shuffles all user *ids* and creates five data folds with distinct user-groups. It ignores the time dimension of the data, but provides user-distinct training and validation folds, which is like the GroupKFold cross-validation approach as implemented in scikit-learn [177]. The *average-user* approach is very similar to the *user-cut* approach. However, each answer of a user is replaced by the *median or mode answer* of this user up to the point in question to reduce within-user-variance. While all the above-mentioned approaches

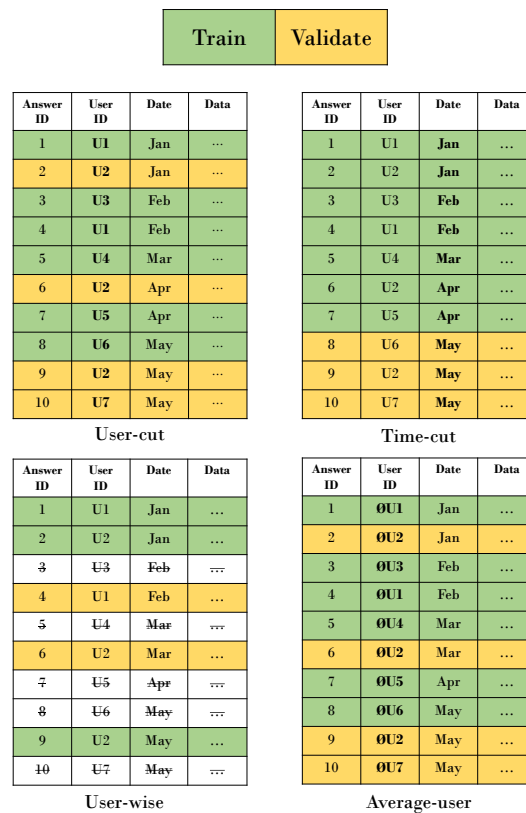


Figure 3.34: Four approaches of data splitting into train folds and validation folds within the train set. Yellow means that this sample is part of the validation fold, green means it is part of a training fold. Crossed out means that the sample has been dropped in that approach because it does not meet the requirements. Users can be sorted by time to accommodate any concept drift.

require only one single model to be trained, the *user-wise* approach requires as many models as distinct users are given in the dataset. Therefore, for each user, 80 % of his or her assessments are used to train a user-specific model, and the remaining 20 % of the time-sorted assessments are used to test the model. This means that for this approach, we can directly evaluate on the test set as each model is user specific and we solved the cold-start problem by training the model on the first assessments of this user. If a user has less than 10 assessments, he or she is not evaluated on that approach.

3.5.3 | Results

We will see in this results section that ignoring users in training leads to an underestimation of the generalizability of the model, the standard deviation is then too small. To

further explain, a model is ranked first in the comparison of all computations if it has the highest final score, and last if it has the lowest final score.

3.5.3.1 | RQ1: What is the variance in performance when using different splitting methods for train and test set?

Considering performance aspects and ignoring the user groups in the data, the time cut approach has on average the best performance on assessment level. As an additional variant, we have sorted users once by time and once by random. When sorting by time, the baseline heuristic with the last known assessment of a user follows at rank 2, whereas with randomly sorted users, the user cut approach takes rank 2. The baseline heuristic with all known assessments on the user-level has the highest standard deviation in ranks, which means that this approach is highly dependent on the use case: For some datasets, it works better, for other it does not. The *user-wise* model approach has also a higher standard deviation in the ranking score, which means that the success of this approach is more use-case specific. As we set the threshold of users to be included into this approach to a minimum of 10 assessments, we have a high chance of a selection bias for the train-test split for users with only a few assessments, which could be a reason for the larger variance in performance. Details for the result are given in Table 3.22.

Could there be a selection bias of users that are sorted and split by time? To answer this, we randomly draw 5 different user test sets for the whole pipeline and compared the approaches' rankings with the variation where users were sorted by time. The approaches' ranking changes by .44, which is less than one rank and can be calculated from Table 3.22. This shows that there is no easily classifiable group of test users.

3.5. 7 observational mHealth studies and 10 years of experience: Can ignoring groups in Machine Learning pipelines lead to overestimation of model performance? Analyses of group-wise Chapter 3. Results validation as well as baseline and concept-drift considerations.

	Users sorted by time		Users split randomly	
	Average rank	Std of average rank	Average rank	Std of average rank
time_cut	2,29	1,50	1,57	0,16
user_cut	3,57	1,72	3,06	0,11
BL user_based last	3,29	1,70	3,46	0,21
average_user	3,86	0,69	3,51	0,36
BL user_based all	3,57	2,37	4,43	0,18
user_wise	4,33	2,07	5,10	0,38
BL assessment_based last	6,86	0,69	6,80	0,12
BL assessment_based all	7,71	0,49	7,66	0,15

Table 3.22: Rank comparison of the four splitting approaches with the four baseline heuristics. Greener means better. Three splitting approaches are on user-level, one is on assessment level. The standard deviation is calculated from the average ranks of 7 datasets. When users are not sorted by time, the approaches are more robust in their rankings, which means that the user cut approach is more likely to work consistently better than the baseline heuristic on user-level. BL = Baseline.

Cross-validation within the train helps to estimate the generalization error of the model for unseen data. On assessment-level, the standard deviations of the weighted F1 score within the train set for all datasets varies between 0.25 % for TrackYourTinnitus and 1.29 % for Corona Health Stress. On user-level, depending on the splitting approach, the standard deviation varies from 1.42 % to 4.69 %. However, on the test set, the estimator of the generalization error (i.e., the standard deviation of the F1 scores of the validation folds within the train set) is too low for all 7 datasets on assessment-level. On user-level, the estimator of the generalization error is too low for 4 out of 7 datasets. We define the estimator of the generalization error as *in range* if its smaller or equals the performance drop between validation and test set. Details for the result are given in Table 3.23.

3.5. 7 observational mHealth studies and 10 years of experience: Can ignoring groups in Machine Learning pipelines lead to overestimation of model performance? Analyses of group-wise Chapter 3. Results validation as well as baseline and concept-drift considerations.

Study	Score	User-Level		Assessment-Level
		user cut	average user	time cut
Corona Check CC	Std Train	1.42%	4.69%	0.54%
	Avg. F1 Train	76.80%	72.50%	77.57%
	F1 Test	67.54%	64.60%	67.98%
	Performance	-9.27%	-7.90%	-9.59%
Corona Health Stress CHS	Std Train	4.95%	3.59%	1.29%
	Avg. F1 Train	54.73%	51.19%	57.47%
	F1 Test	51.32%	53.10%	53.15%
	Performance	-3.41%	1.91%	-4.31%
Corona Health Mental Health Adolescents CHY	Std Train	1.80%	1.46%	0.71%
	Avg. F1 Train	98.85%	98.28%	98.89%
	F1 Test	97.63%	94.86%	98.18%
	Performance	-1.21%	-3.42%	-0.71%
Corona Health Mental Health Adults CHA	Std Train	3.75%	3.79%	1.23%
	Avg. F1 Train	65.80%	66.73%	69.87%
	F1 Test	61.79%	62.24%	63.05%
	Performance	-4.00%	-4.48%	-6.81%
Corona Health Physical Health Adults CHP	Std Train	2.24%	2.47%	1.06%
	Avg. F1 Train	47.79%	43.70%	53.25%
	F1 Test	45.38%	46.53%	45.97%
	Performance	-2.40%	2.84%	-7.28%
Track Your Tinnitus TYT	Std Train	2.25%	3.78%	0.25%
	Avg. F1 Train	54.88%	45.97%	58.70%
	F1 Test	56.26%	40.57%	57.26%
	Performance	1.38%	-5.40%	-1.44%
Unification of Treatments and	Std Train	2.51%	1.62%	0.36%

Table 3.23: Performance scores and standard deviations of the seven use cases on user- and assessment-level. For the user-level, there are two splitting approaches shown: *User-cut*, with users sorted by time of sign-up, and *average-user*, where an answer given by a specific user is averaged with the users' previously given answers. Red numbers indicate the performance drop from train to test. f_1^{train} conforms the average f_1 scores of the validation folds of the train set.

Both approaches, user- and assessment, overestimate the performance of the model during training. However, the quality of estimator of the generalization error increases if users are split on user-level.

3.5.3.2 | RQ2: In which cases is the development, deployment and maintenance of a ML model compared to a simple baseline heuristic worthwhile?

For our 7 datasets, the baseline heuristics on a user-level perform better than those on assessment-level. For the datasets *Corona Check (CC)*, *Corona Health Stress (CH)*, *TrackY-*

ourTinnitus (TYT) and UNITI, the last known user assessment is the best predictor within the baseline heuristics. For the psychological Corona Health study with adolescents (CHY) and adults (CHA), and physical health for adults (CHP), the average of the historic assessments is the best baseline predictor. The last known assessment on an assessment-level as a baseline heuristic performs worse for each dataset compared to the assessment level. The average of all so far known assessment as a predictor for the next assessment-independent from the user - has worst performance within the baseline heuristics for all datasets except CHA. Notably, the larger the number of assessments, the more the all-instances-approach on assessment-level converts to the mean of the target, which has high bias and minimum variance.

		CC	CHS	CHY	CHP	CHA	TYT	UNITI
User-Level	Last instance	0.604 (0.008)	0.567 (0.008)	0.626 (0.028)	0.580 (0.014)	0.671 (0.008)	0.250 (0.003)	0.515 (0.005)
	All instances	0.555 (0.008)	0.558 (0.016)	0.687 (0.037)	0.660 (0.020)	0.698 (0.006)	0.190 (0.003)	0.504 (0.005)
Assessment-Level	Last instance	0.445 (0.008)	0.273 (0.004)	0.288 (0.040)	0.275 (0.012)	0.313 (0.013)	0.205 (0.003)	0.254 (0.007)
	All instances	0.302 (0.006)	0.138 (0.018)	0.233 (0.022)	0.176 (0.018)	0.317 (0.011)	0.187 (0.003)	0.173 (0.011)

Table 3.24: Results of the four baseline approaches on the 7 datasets. The first number of a cell is the average f_1 score with the standard deviation (std) in brackets: $f_1(std)$. For each dataset, the top score is marked green while the lowest score is marked orange.

These results lead us to conclude that recognizing user groups in datasets leads to an improved baseline when trying to predict future ones from historical assessments. When these non-machine-learning baseline heuristics are then compared to machine learning models without hyperparameter tuning, it is found that they sometimes outperform or similarly outperform the machine learning model.

Kind of model	ML	Baseline	Baseline	ML	ML	ML	Baseline	Baseline
Approach name	Time Cut	User-Level Last	User-Level All instances	User Cut	Average User	User Wise	Assessment-Level	Assessment-Level
Average rank	2.29	3.29	3.57	3.57	3.86	4.33	6.86	7.71
Std average rank	1.50	1.70	2.37	1.72	0.69	2.07	0.69	0.49

Table 3.25: Average rank of the approach for all datasets, including the standard deviation of the rank one line below. On average, the baseline heuristics on the user-level are ranked slightly better than the ML model on a user-level. Best rank is left, worst rank is right.

The approaches ranking in Table 3.25 shows the general overestimation of the performance of the *time-cut* approach as this approach is ranked best on average. It can be also seen that these approaches are ranked closely to each other. Because we only subtract 0.5 of the standard deviation of the f_1 scores of the validation folds, approaches with a higher standard deviation are less punished. This means, in turn, that the overestimation

of the performance of the splits on assessment-level would be higher if the pre-factor of σ was higher. Another reason for the similarity of the approaches is that the same model architecture has been finally trained on all assessments of all train users to be evaluated on the test set. Thus, the only difference of the rankings results from the standard deviation of the f_1 scores of the validation folds.

To answer the question whether it is worthwhile to turn a prediction task into an ML project, further constraints should be considered. The above analysis shows that the baseline heuristics are competitive to the non-tuned random forest with much lower complexity. At the same time, the overall results are an f1 score between 55 and 65 for a multi-class classification with potential for improvement. Thus, the question should be additionally asked, from which f_1 score can be deployed, which depends on the use case, and in addition it is not clear whether the ML approach can be significantly improved by a different model or the right tuning.

3.5.4 | Discussion

The present work compared the performance of a tree-based ensemble method if the split of the data happens on two different levels: User and assessment. It further compared this performance to non-ML approaches that uses simple heuristics to also predict the target on a user- or assessment level. We quickly summarize the findings and then discuss them in more detail in the sections below.

- Ignoring users in datasets during cross-validation leads to an overestimation of the model's performance and robustness.
- For some use cases, simple heuristics are as good as complicated tree-based ensemble methods. Within this domain, heuristics are more advantageous if they are trained or applied at the user level. ML models also work at the assessment level.
- Sorting users can simulate concept drift in training if the time span of data collection is large enough. The results in the test set change due to shuffling of users.

3.5.4.1 | Limitations of our findings

The - still - small number of 7 use cases itself has a risk of selection bias in the data, features, or variables. This limits the generalizability of the statements. However, it is also arguable whether the trends found turn in a different direction when more use cases are included in the analysis. We do not believe that the tendencies would turn. We restricted the ML model to be a random forest classifier with a default hyperparameter

set up to increase the degree of comparability between use cases. We are aware that each use case is different and direct comparability is not possible. Furthermore, we could have additionally evaluated the entire pipeline on other ML models that are not tree-based. However, this would have added another dimension to the comparison and further complicated the comparison of the results. Therefore, we cannot preclude that the results would have been substantially different for non-tree-based methods, which can be investigated further in future analyses.

Future research of this user-vs.-assessment-level comparison could include a hyperparameter tuning of the model on each use case, a change of model kind (i.e., from a random forest to a support vector machine) to see whether this changes the ranking. The overarching goal remains to obtain the most accurate estimate of the model's performance after deployment.

3.5.4.2 | Baseline heuristics

We cannot give a final answer to what can be chosen as a common baseline heuristic. In machine learning projects, a majority vote is typically used for classification tasks, and a simple model such as a linear regression can be used for regression tasks. These approaches can also be called naive approaches since they often do not do justice to the complexity of the use case. Nevertheless, the power of a simple non-ML heuristic should not be underestimated. If only a few percentage points more performance can be achieved by the maintenance- and development-intensive ML approach, it is worth considering whether the application of a simple heuristic such as "the next assessment will be the same as the last one" is sufficient for a use case. Notably, Cawley and Talbot argue that it might be easier to build domain expert knowledge into hierarchical models, which could also function as a baseline heuristic [317].

3.5.4.3 | The impact of shuffled users

To retain consistency and reproducibility, we kept the users sorted by sign-up date to draw train and test users. The advantage of sorting the users is that one can simulate potential concept drift during training. The disadvantage, however, is an inherent risk of a selection bias towards users that signed up earlier for a study. From Table 3.22, we can see that the overfitting of users increases when we shuffle them. We conclude this from the fact that the difference between the average ranks of the approaches *time cut* and *user cut* increases. The advantage of shuffling users is that the splitting methods seem to

depend less on the dataset. This can be deduced from the reduced standard deviation of the ranks compared to the sorted users.

3.5.4.4 | Performance drops from validation folds to test set

Regardless of the level of splitting (user- or assessment-level), one can expect a performance drop if unknown users with unknown assessments are withheld from the model in the test set. When splitting at the user-level, the performance drop is lower during training and validation compared to the assessment-level. However, it remains questionable why we see this performance drop in the test set at all, because both, the validation folds and the test set contain unknown users with unknown assessments. A possible cause could be simple overfitting of the training data with the large random forest classifier and its 100 trees. But, also a single tree with max depth = number of features and balanced class weights has this performance drop from the validation to the test set. One explanation for the defiant performance drop could be that during cross-validation information leaks from training folds to validation folds, but not to the test set.

3.5.4.5 | Final thoughts and recommendations

A simple heuristic is not always trivial to beat by an ML model, depending on the use case and the complexity of the search space. Thinking of the complexity that a ML model adds to a project, a heuristic might be a valuable start to see how well the model fits into the workflow and improves the outcome. A frequent communication with the domain expert of the use case helps to set up a heuristic as a baseline heuristic. In a second step, it can be evaluated whether the performance gain from an ML model justifies the additional development effort.

3.5.5 | Data and Code availability statement

According to the General Data Protection Regulation of the European Union, the data to replicate these results are available upon request to the corresponding author.

All code to replicate the results, models, numbers, figures, and tables is publicly available to anyone on <https://github.com/joa24jm/UsAs>. Any supplementary material is also in this repository.

3.5.6 | Acknowledgements

Ethics Approval for the UNITI-RCT and the App was obtained by the local ethics committees at all investigator clinical sites. The Track Your Tinnitus (TYT) study was approved by the Ethics Committee of the University Clinic of Regensburg (ethical approval No. 15-101-0204). The Corona Check (CH) study was approved by the Ethics Committee of the University of Würzburg (ethical approval no. 71/20-me) and the university's data protection officer and was carried out in accordance with the General Data Protection Regulations of the European Union. The procedures used in the Corona Health (CH) study were in accordance with the 1964 Helsinki declaration and its later amendments and was approved by the ethics committee of the University of Würzburg, Germany (No. 130/20-me).

Funding This work was partly funded by the ESIT (European School for Interdisciplinary Tinnitus Research [328]) project, which is financed by European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement number 722046 and the UNITI (Unification of Treatments and Interventions for Tinnitus Patients) project financed by the European Union's Horizon 2020 Research and Innovation Programme, Grant Agreement Number 848261 [323]. J.A. and R.P. are supported by grants in the projects COMPASS and NAPKON. The COMPASS and NAPKON projects are part of the German COVID-19 Research Network of University Medicine ("Netzwerk Universitätsmedizin"), funded by the German Federal Ministry of Education and Research (funding reference 01KX2021). This publication was supported by the Open Access Publication Fund of the University of Wuerzburg.

Informed Consent We have obtained informed consent from all participants. We have complied with all relevant ethical regulations.

Author contributions statement J.A. primarily wrote this paper, created the figures, tables and plots, and trained the machine learning algorithms. R.P. supervised and revised the paper.

Conflict of interest statement The authors declare no competing interests.

Discussion

After having presented the research papers subsequently and presented the results, we would like to summarize the overall findings, and discuss limitations. In doing so, we divided this chapter into five sections. The first section 4.1, labeled "Recap of the research questions", revisits the research questions introduced in previous sections. This section establishes connections between the research questions articulated in the introduction, the thesis-contributing papers, and provides a concise overview of the principal findings highlighted in Chapter 3. Following this, the subsequent section 4.2, titled "Overall interpretation of results", offers a comprehensive analysis of the cumulative research outcomes. The goal is to present a holistic perspective that highlights the broader implications and significance of the research findings. Within the third section 4.3, "Discussion of Limitations", the study's inherent constraints are openly acknowledged. These limitations are categorized into those arising from the employed models, limitations associated with the utilized data sources, and constraints inherent to the specific domain of study. The fourth section 4.4, "Future Research", suggests avenues for possible future investigation and points to directions for further scientific exploration. The intent is to guide forthcoming research endeavors in further elucidating the addressed research directions. Lastly, the concluding section 4.5, "Concluding Summary", encapsulates the synthesized insights gleaned from the entirety of the thesis. Serving as the conclusion, this section concisely recaps the key takeaways and underscores the broader significance of the undertaken research.

4.1 | Recap of the research questions

The mHealth domain is a specific subdomain in medicine. As a relatively new field, it also brings the additional challenge of communication at points of intersection be-

tween medicine and informatics. More particular, physicians and data scientists or ML engineers have to discuss and to understand each other's challenges and problem understandings. In doing so, they need a common language. With this background given, one can derive the three research questions that are discussed in this thesis and also shown in Table 4.1.

Main RQ	Sub Research Question	Finding	Section
How can machine learning help confirming or broaden domain knowledge within mHealth data?	Can we predict the gender of a user using EMA data?	Gender can be predicted with an accuracy of 81.7 %.	3.1
	Which ML architecture is most suitable for the gender prediction task?	For tabular data, tree-based models have the highest prediction power. In this case, a random forest classifier.	3.1
	Can momentary tinnitus and tinnitus loudness be predicted using EMA data?	Both targets can be predicted well beyond guessing with 94 % accuracy and 7.9 % mean absolute error.	3.2
	How can the generalizability of the model be ensured?*	Better cross-validation techniques lead to more precise error estimates.	3.5
How can one reach explainability in the presence of mHealth data when using Machine Learning?	Which are the features with the highest importance to predict gender of users of the TYT app?	Most contributing features come from the reported worst symptom.	3.1
	Using multi-modal data, which country- and season specific differences are between tinnitus patients?	Using partial dependence plots among others, tinnitus and its symptoms vary between seasons and countries.	3.2
	Which XAI methods are mostly used for medical use cases?	Methods in Python that are easily accessible and widely usable: SHAP, LIME among others.	3.4
	From which data type is the input fed to a model?	The distribution is: Text (65 %), image (32 %), text (3 %), and audio (0 %).	3.4
	How many XAI methods can be potentially understood by patients?	Our estimation: Only 16 % of the methods reported in the included papers.	3.4
Which guidelines can be beneficial for the use of ML within the mHealth domain?	Is the source code provided and is data publicly available?	On average, 21 % of the paper make code and 27 % make data available with stagnant trends.	3.4
	How well is the ML pipeline described?	On average, 2.15 out of 3 with a positive trend to better descriptions.	3.4
	How does ML performance vary using different train-test-split methods?	Split on assessment level leads to an overestimation of model performance.	3.5
	When does the deployment of a complex ML model add value compared to a simple heuristic?	The choice of ML model or simple heuristics depends on the use case and external factors.	3.5
**	Are there differences in the distribution of reported symptoms between countries, sex, or age groups?	No statistically significant difference between countries, age groups or sex for reported symptoms.	3.3
	Does app usage correlate with the number of reported Covid infections?*	The app provided easily accessible information on corona symptoms, but app usage did not seem to correlate with Covid-19 infections.	3.3
	How well is the offer of a free Corona Check app with additional information accepted?*	mHealth apps are valuable tools for multimodal and longitudinal data collection.	3.3

Table 4.1: Summarizing of the research questions and findings addressed in this dissertation. The table also links sub-research questions to one of the three main research questions as described in the above section.

*This question is not explicitly stated, but indirectly addressed in the corresponding papers.

**The research questions of the Corona Check project are not directly related to the main research questions of this thesis as the Corona Check paper does not include any ML. The contribution to the software development through the Excel Loop with the creation of the codebook led to the data that was evaluated in this thesis, which is why the paper is part of the cumulative dissertation. The research questions from the paper are listed here for completeness.

Within the mHealth domain, *how can ML help confirming or broaden domain knowledge within mHealth data?* (Main RQ 1). The second research question is derived from the first: *How can one reach explainability in the presence of mHealth data when using ML?* (Main RQ 2). And research question 3 asks, *which guidelines can be beneficial for the use of ML within the mHealth domain*, which we refer to as Main RQ 3. The contributing papers of the main section then address the main research question by sub-research questions, which are briefly summarized in Table 4.1.

Within **Main RQ 1**, (*How can machine learning help confirming or broaden domain knowledge within mHealth data?*), we show that in a balanced test set (*balanced*: uniformly distributed target labels in the test set), the gender of a TYT user can be predicted with an accuracy of 81.7 % for our given features, which is well beyond guessing, which would be 50 % as the test set is balanced in this binary prediction task. By comparing the performance and speed of different algorithm architectures, we also show that for tabular data, tree-based ensembles models such as a random forest or gradient boosting machine still provide a reasonable trade-off for speed and performance. They are comprehensible, fast to train, and perform - within tabular data - only slightly worse than neural networks, which are bigger in size and more difficult to comprehend.

In **Main RQ 2**, (*How can one reach explainability in the presence of mHealth data when using Machine Learning?*), we then ask how explainability can be reached when using mHealth data in ML problems. Use cases on the TYT dataset aimed to predict gender or tinnitus perception using large ensembles of decision trees. These ensembles were explained by enhancing feature importance methods (i.e., the random forest feature importance) and partial dependence plots. Using these explainability methods, we then concluded, together with domain experts, that reported epiphenomenons like stress or insomnia are valuable features to predict gender. For tinnitus perception, the perceived perception varies for different temperatures, seasons and it varies between countries. In showing that gender can be meaningfully predicted, we confirmed the hypothesis of the tinnitus research community that tinnitus varies based on gender using a new method from the ML domain. In showing that we can predict tinnitus perception (section 3.2), we broadened the tinnitus domain knowledge by showing that **season and temperature** are valuable features and correlate with tinnitus perception. In other words, the non-parametric ensemble model showed a correlation of input *A* (season, country) and output *B* (tinnitus perception). For these papers in sections 3.1 and 3.2, we mainly used partial dependence plots, statistics and feature importances to explain the models. On the literature review, which was carried out in section 3.4, we then found that the most common explainability methods in the medical are SHAP, LIME, and Grad-Cam. We did not find a single paper that matched our inclusion criteria of using **audio data** within

a medical ML use case among with an explainability method. Still, most of the papers used tabular or image data for the ML input, with a slight tendency towards more image data. We then further estimate that only 16 % of the methods reported could be potentially understood by patients, which are an important stakeholder group for the implementation of ML systems in the medical domain.

Main RQ 3, (*Which guidelines can be beneficial for the use of ML within the mHealth domain?*) is then partly addressed in section 3.4, in which we asked about the code and data sharing ratios of the ML explainability medical papers that have been included in the literature review. On average, about one fifth (21 %) of the papers publish their code and one quarter (27 %) publish their data. We also obtained a positive trend in the precision of the description of ML pipelines to reproduce the reported results. We calculated an average score of 2.15 on an ordinal scale from 1 to 3 where larger means better (section 3.4). We then asked in section 3.5, how ML performance varies, given different train-test-split techniques. We found two things: First, ignoring user groups in the data leads to an overestimation of the external validity of the model. In other words, the model overfits on strongly represented users (synonym: power users) and learns user patterns instead of relevant features. Second, a simple heuristic ("The next value equals the last one") sometimes outperform complex ML ensembles in both variance and performance. We then state that the choice between a ML model and a simple heuristic depends on the use case and external project contributing factors. To summarize, from what we saw in the literature review in section 3.4, and the cross-validation evaluation in section 3.5, we would like to suggest the following guidelines, which finally answers Main RQ3:

- During data cleansing, the data understanding within CRISP-DM should not only be validated with subject matter experts, but also with software engineers, who are aware of how the data is technically received and stored to prevent rounding or operating system related errors.
- The report of the ML pipeline should not only include hyperparameters of the model used, such as the detailed architecture of a neural network or random forest, but also the pre-processing, standardization, normalization, and data cleansing steps that were necessary to arrive from raw data to a ML-ready dataset.
- Code and data should be easily accessible, and shared with documentation. If data must not be shared publicly, it should always be available upon reasonable request for research and reproducibility.
- The data must have a codebook that contains at least variable name, variable meaning, coding, coding meaning and optionally, a comment and unit section.

- The data collection process should be briefly documented. Who collected when, where and with which purpose the data? This information could be beneficial later to explain noise in the data or draw limitations.
- The overall use of a ML algorithm should be evaluated against a context-aware baseline model, which could be a simple heuristic such as "the next value is the last one" or the mean of the target in the train set.
- Hidden user groups should be detected and separated from each other during training to improve external validity of the model.
- Model explainability can be enhanced and used to improve model performance and deepen stakeholder trust.

As in Table 4.1 shown, the research questions of the **Corona Check paper** in section 3.3 cannot be meaningfully linked to one of the main research questions stated above. However, due to my contribution to the data building pipeline and my analyses, section 3.3 is part of this thesis and this section summarizes all results. So, we found that there are no statistically significant differences between countries, age groups or sex for self-reported Covid-19 symptoms, which means that these users have experienced the coronavirus in a comparable way during this time based on their symptom distribution. We also did not find a correlation between the overall app usage and the world-wide Covid-19 development in 2021. However, this paper shows that mHealth apps are valuable tools for multi-modal and longitudinal data collection, which, in turn, can be used to potentially feed ML algorithms. By developing and launching this app in an early stage of the pandemic, we helped relieve overloaded hotlines by giving people information about the coronavirus through the app, and by filling out the questionnaire, they got feedback on whether they should see a doctor or isolate themselves.

4.1.1 | What is the medical contribution?

In this subsection, we would like to succinctly point out the **medical contribution** of this thesis. In doing so, we distinguish between a linking contribution, which shows the relevance of bridging the gap between medicine and informatics, and a domain-specific contribution for thesis-contributing papers that required subject matter experts from both the tinnitus domain, and severe acute respiratory syndrome coronavirus type 2 domain. The whole medical contribution is summarized in Figure 4.2. On an abstract level, this thesis aims to strengthen the field of **Digital Medicine** by developing a deeper data understanding for mHealth related data science projects and reporting common

pitfalls in data collection and processing. The abstract level contribution is sketched in Figure 4.1. By analyzing, aggregating, processing, cleaning, and interpreting longitudinal

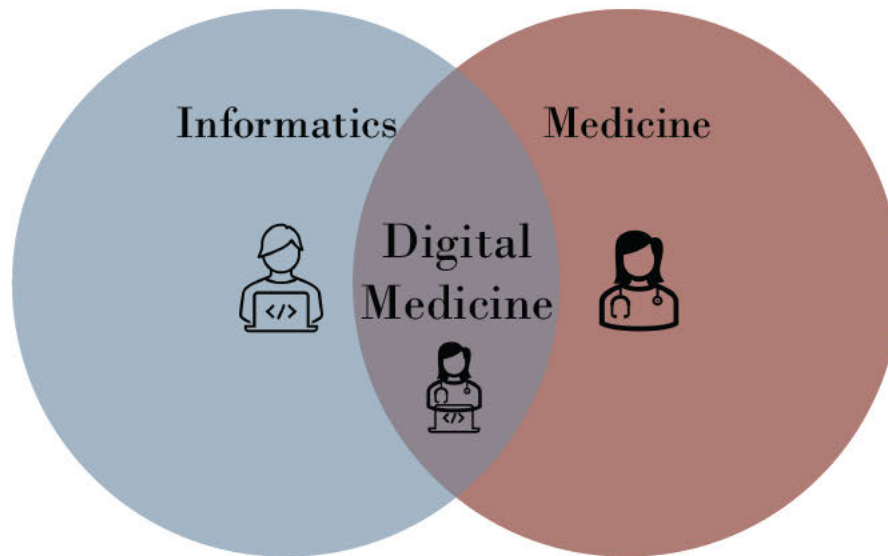


Figure 4.1: The development of a deeper data understanding and common pitfalls in the medical domain is one of the medical contributions of this thesis. It aims to strengthen

Confirm and extend domain knowledge If a model can predict an output B with a given input A , we can say that A is associated with B . We then can confirm known domain-specific associations with subject matter experts. Further, if the model is highly generalizable and robust, we can explore unknown associations from A to B within a CRISP-DM feedback loop or a **subject matter expert in the loop**. We have chosen seven different multi-model and longitudinal datasets with the hope to confirm beknown or find new A to B mappings. As medical domains, we chose the gender aspect in tinnitus research from section 3.1, and the season and temperature aspect from section 3.2. For the tinnitus gender study (section 3.1), we found that there exist gender-related differences in coexistent symptoms for tinnitus patients. The analysis suggest women have more sleep onset difficulties, while men struggle to sustain attentive engagement in conversations. For the tinnitus country study (section 3.2), our analysis indicates that tinnitus does not typically increase in winter months, but is influenced by temperature and season, with a potential peak during the summer. Regarding the Corona Check app, it provided

information on the coronavirus, and thus supported overburdened coronavirus hotlines. We further found that the virus was reported equally by users of different countries, gender and age groups - we did not find statistically significant different distributions between these groups which suggests people experienced the infection equally regardless country, gender, or age.

Bridging the gap between medicine and informatics The CRISP-DM cycle from subsection 2.3.1 explains the feedback loop that should be carried out between subject matter experts and the data analysts in a ML project. This feedback loop requires a basic mutual understanding of each other's domains as well as a common language that clarifies synonyms and homonyms, which might hinder effective interdisciplinary communication. The literature review from section 3.4 aims to do that: Strengthening the bridge between medicine and informatics. It is a paper that is primarily addressed to physicians, written by computer scientists. It captures and explains the state of the art of ML explainability, provides a taxonomy that helps to overcome homonyms and synonyms, and it explains limitations as well as possible applications of the provided XAI methods.

The user-assessment paper from section 3.5 highlights a potential importance of interdisciplinary communication between physicians and data scientists. By close communication, subject matter experts can help data scientists to detect hidden groups in data. However, this has been studied in more detail in only two papers (section 3.4 and 3.5) so far and needs more research for sharper limitations and more precise conclusions. This information may then help to provide better estimates for model performance in out-of-distribution data. In addition, the subject matter expert can also help in building simple heuristics, which can be evaluated against a complex ML approach and thus help to assess the effort of training and deploying a complex model. Again, this requires interdisciplinary, and open communication with a common language. The literature review (section 3.4) helps to find this common language.

4.1.2 | What is the informatics contribution?

The contribution for both medicine and informatics is partly mentioned in Table 4.1. However, for the sake of clarity, we summarize the contribution of for the informatics field in this subsection.

- This thesis developed a deeper understanding of implementation challenges of ML and XAI on mHealth data.

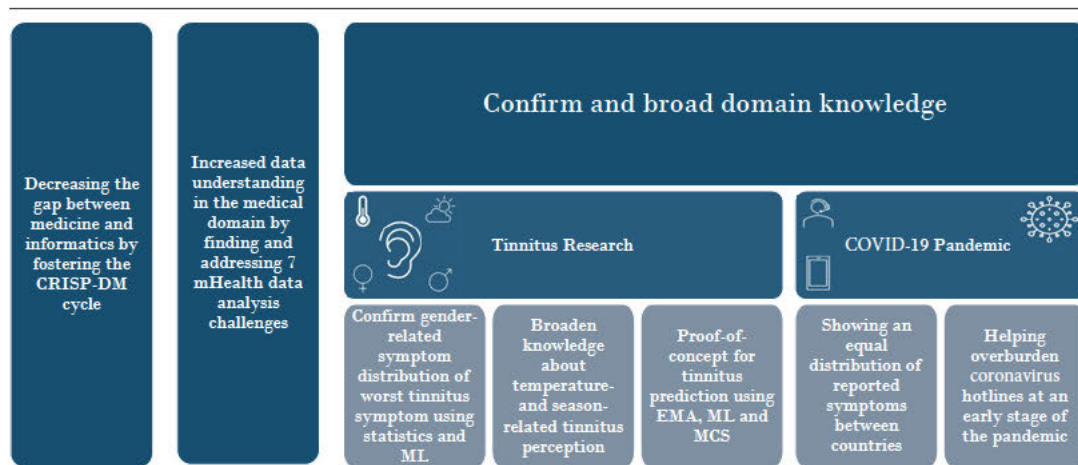


Figure 4.2: Summarization of the medical contribution of this thesis by topic and level.

- It also questions whether ML is a tool of choice at all. The answer is, it depends on the use case, the target group, the workflow of the use case, and the environment where the tool shall be implemented.
- It further makes considerations for baseline models in different papers: Dummy models, random guesses, majority votes, simple domain-related heuristics, or benchmark models.
- It provides alternative train-test split approaches for data containing groups, taking time into account.
- It summarizes common explainability methods, and their applications as well as limitations in the medical domain.
- It confirms that for tabular data, tree-based ensembles provide a reasonable trade-off for performance and model-size, compared to neural networks or support vector machines.
- It advises to correct for user groups in cross-validation leads to obtain more precise out-of-distribution error estimates of the model.
- It suggests to sort data by time before train-test splits to simulate concept drift for longitudinal data sets, which can create more realistic test sets.
- There is a list of guidelines for mHealth related ML projects that help to address the 7 mHealth challenges to Figure 1.1.

- Another finding: SHAP, Grad-Cam and intrinsic interpretable methods are most used tools for XAI. We think they are widely adopted, because they are easy to use (using Python's `pip install`), post-hoc and model-agnostic.
- Audio data is still rarely used for medical use cases with XAI requirements.
- We estimate that maximum 1 of 5 XAI methods can be understood by end users in medicine (i.e., patients).
- Ignoring groups in training data leads to overfitting and overestimation of model performance.
- Simple heuristics ("The next value equals the last one") sometimes beat complex ML models.

4.2 | Overall interpretation of results

Main RQ 1 (*How can ML help confirming or broaden domain knowledge within mHealth data?*) is addressed in sections 3.1, 3.2 and 3.5. All of these three papers apply tree-based ML algorithms on multi-modal mHealth data to detect *A-to-B* mappings in a multi dimensional search space that is too complex for the human brain to comprehend. We show in section 3.1 that tree-based models are still a valuable option among a large variety of different ML architectures to address prediction tasks for tabular data. This indeed answers the sub-research question *Which ML architecture is most suitable for the gender prediction task*. By applying ML algorithms on multiple use cases within different mHealth EMA datasets, we first show that these algorithms have higher performance scores (such as F1, and Mean Absolute Error) than guessing or simple heuristics, and further, they are able to learn *A-to-B* mappings of arbitrary complexity. This then answers Main RQ 1. Machine Learning can indeed help to confirm or broaden domain knowledge if

- the model performs significantly better than heuristics or random guessing,
- the subject matter expert is involved in a feedback loop using CRISP-DM while developing the model and analyzing the training data,
- challenges deriving from multi-modal data are appropriately addressed, and
- hidden user groups are separated from each other during training for validation.

Regarding Main RQ 2 (*How can one reach explainability in the presence of mHealth data when using Machine Learning?*), we would argue from the results of section 3.2 that a step-wise increase of the complexity - even when using explainability methods - helps to comprehend and understand *why* a model makes a certain prediction. Rather less complex interpretability methods are decision trees (as post-hoc explainability) or partial dependence plots. If a partial dependence plot, i.e., shows that insomnia or sex are valuable features to decrease entropy in the target distribution, then SHAP or LIME can be used to verify or disprove this hypothesis in a multi-dimensional space while using a completely different approach. Or similarly, if SHAP strongly suggests that insomnia is an important feature, then a random forest feature importance should state something similar. Different approaches are indeed another step to reach explainability. When applying different explainability methods on the same model, i.e., to determine feature importance, different methods may yield to slightly different results (see section 3.1, table 3.6). While this can be explained by the model architecture and the approach of the explainability method, the most important key take-away from Table 3.6 is that all of methods used provide a **tendency** which feature is rather less or more important.

The last factor that contributes to explainability is the domain expert. According to CRISP-DM, there should be an agile, open communication and knowledge transfer between data scientist and domain expert. That is, if the ML engineer or data scientist found a correlation between input *A* and output *B*, they should verify this in a discussion with the domain expert. We would like to call this approach the **domain expert in the loop**. To finally and briefly answer Main RQ 3: Explainability in the presence of mHealth data when using ML can be reached if:

- The complexity of the explanatory methods is gradually increased with concurrent mutual verification of each others results,
- different explainability methods rank the same feature similarly important, and
- the **domain expert in the loop** should communicate agilely and openly with the data scientist (and vice versa) to verify results around the development process.

*Which guidelines can be beneficial for the use of ML within the mHealth domain, which is addressed in Main RQ 3, is a question that cannot be fully answered in one paragraph. The publications in the field of AI increase rapidly due to the large research community and various branches such as computer vision, language models, and tabular or time series data. In the following, we would like to summarize the guidelines that **generally** apply for ML projects and add guidelines that apply for mHealth ML projects particularly. **General guidelines include:***

- Ensure data quality together with domain experts [329]
- Select given features, engineer new features or choose an end-to-end approach using Deep Learning [330]
- Design the model architecture and choose metrics that align with the project's desired outcome [331]
- Ensure model explainability for stakeholders such as ML engineers, physicians, patients, users, authorities [287]
- Consider ethical implications of the data, potential biases and impact on stakeholders [332]
- Ensure alignment with relevant regulations such as the EU AI Act, Software as a medical device regulations, and the general data protection regulation.

MHealth ML project guidelines may include:

- Correct your dataset for power users which may induce bias.
- Ensure equidistant measurements by reminding users to fill out recurring assessments - provide incentives to keep users using the app.
- Think of the time axis of longitudinal studies and correct for concept drift.
- Avoid optional questions in the assessments to avoid missing value treatment. If you have missing values, you may preferably use user-wise imputation methods.
- Correct for operating system specific measuring inaccuracies such as rounding or cut-off errors. Consult app and backend developer for to find other potential data issues that are not domain related.
- Ensure user identity by providing secure logins to the study and remind users to not pass their phone to other people when filling out the assessment.
- Make the model complex enough to capture the heterogeneity of all users, but simple enough to estimate out of distribution users (also known as the bias-variance trade-off).

In this section, we answered the three main research questions briefly and provided an overall interpretation of the results. The following section points out the limitation of these results.

4.3 | Limitations of this work

This section is subdivided into three subsections.

Model-related, data-related and domain-related limitations. Model-related limitations discuss potential performance decrease due to overfitting and out-of-distribution data. Data-related limitations refer to the challenges that naturally arise when working with mHealth MCS data, such as unknown user identity or power users. In domain-related limitations, we discuss the restrictions of the medicine-related results.

4.3.1 | Model-related limitations

Regarding sections 3.1 (Tinnitus Gender) and 3.2 (Tinnitus Country), we could not evaluate the models on out-of-distribution data as they never got deployed. From ML Operations, we know that there are many challenges when deploying a model and the ML code itself is just a small part of the project. Some of these issues are concept drift, data shift, and overfitting. That is, the reported performances of the models in both papers may decrease after deployment. This, in turn, can lead to a re-engineering of the features and thus slightly change the outcome regarding feature rankings and feature importances. Another limitation that the models of these papers have in common: They have been evaluated on assessment-level. From section 3.5, we know that this can lead to overfitting and an overestimation of model performance. In particular, overfitting means that the model might have learned filling out behaviour of **power users** (not desired) instead of domain-relevant feature patterns (desired). From the introduction, we also know that **user identity** is an issue. We never really know if a user passes his or her mobile device to another person, asking to fill out the current assessment. This leads to noise in the data, which limits our findings.

4.3.2 | Data-related limitations

In the introduction, we have addressed the problem of **non-equidistant measurements**. From the analysis of section 3.5 (Figure 1.4), we know that there is a large variance of the time gap of two filled out assessments of a user. If we then train a model to predict the next assessment and the study design suggests a time gap of two weeks, the model might predict a time gap of 18 or 22 days because of irregular filling out behaviour, which is reflected in the train data. In section 3.5, we used different datasets to better estimate the out-of-sample performance of our ML models after deployment. The datasets that we used, however, might be limited in comparability because of the

different user behavior in mHealth studies. The user behavior between studies can vary because of time-related concept drift (early data was collected in 2014, other data was collected during the pandemic in 2020). The user behavior can further vary because of the different circumstances, in which these users fill out the assessment. A user with severe tinnitus symptoms has a different motivation to complete the questionnaire than a disinterested user from the Corona Health study during lockdown. **Missing values** remain a problem for assessments with non-required, optional questions. Although there exist many techniques for missing value treatment, such as user-based missing value treatment or *k*-nearest-neighbor, these imputed values are *estimates* of non-existent values and thus add noise to the train data. **Operating system specific measuring inaccuracies** are one issue that we detected during our analyses. There might be more non-detected issues that add noise or inaccuracies to the train data and thus limit the results, worsen model robustness, or model performance.

The literature review from section 3.4 contains data until March 7 in 2022. The AI research field, however, is an exponentially growing field and since 2022-03-07, there have been published many more papers, which could lead to different rankings other than we found. Also, even within the search time from 2008 to 2022, we could not cover all available data bases for medical AI literature as we were limited in people to carry out the review. However, we are confident that with over 2500 abstracts, we have a representative sample from the literature to have reliable estimators of the true values. A general characteristic of our mHealth studies is that they are **convenience samples** with low barriers to entry and exit. For example, with the Corona Health and Corona Check apps, anyone can download the app, sign up for the studies, and fill out assessments. That leads to a risk of a selection bias in multiple directions. The Corona Check app, i.e., was mostly filled out by Germans, followed by users from India and South Africa (Figure 1.3). Thus, when we say that we did not find differences in the distributions of reported symptoms between countries, we are referring primarily to these countries. If we then look at the age distribution, we see that the average of our app users is not the same as the average of the population: mainly younger people use these apps. The low exit barriers then lead to an exponential drop out of the app users, which then leads to the challenge of **power users**, which we described in paragraph 1.2.1.

4.3.3 | Domain-specific limitations

Regarding the tinnitus domain from section 3.2, observing significant seasonal and geographical variations in *momentary tinnitus*, it's essential to note that these differences **do not imply causality** between features and the target. Although our findings offer

valuable insights for tinnitus research, limitations should be acknowledged. Contributing features beyond those observed, such as air pressure, stress, and sunlight exposure, could contribute to tinnitus variability. Varying user counts among countries could further introduce selection bias. Additionally, individual experiences of tinnitus may differ from the identified trends. Thus, the applicability of these findings to individuals is restricted.

Regarding the coronavirus domain, the virus is known to have mutated continuously during our data collection phase and there were different variants (including the well-known ones like Omicron O , and Delta Δ). The symptom distribution also varies among the coronavirus variants, which we did not consider in the evaluation, because we did not know whether the users were infected and if so, with which variant. The data were not granular enough for this.

Also, one of our hypotheses from the Corona Check paper is that the app helped take pressure off coronavirus hotlines because more people were getting information from the app instead of calling the hotline. However, we were never able to investigate whether and to what extent our app actually relieved hotlines. Nevertheless, we can cautiously and logically conclude that the number of people calling a hotline will fall if more people find out about an app with the same information content.

After giving model-, data-, and domain-specific limitations, we point out future research directions in the next section.

4.4 | Future research

Because this work is placed at many intersections, there are many direction for further research directions. We would therefore like to make suggestions here with an excerpt that we consider promising.

Addressing the exponential user drop off Most of the user stop using the app after one or two assessments. This stops us from creating user-wise longitudinal datasets, which would be valuable in training robust models. **Gamification** could be one approach to keep users using the app. But how promising is that? Which other approaches exist and how effective are they? Another idea is to provide **interaction using large language models** (LLM) such as Generative pre-trained transformer 4 (GPT-4) [333] or an app-based social network, which allows users to get in contact with each other to exchange about their disease experiences.

Deepen explainability knowledge for large language models Some researchers consider the new generation of LLMs as artificial general intelligence (AGI) [79]: An AI system that can solve tasks it was not explicitly trained for. These systems are also able to explain their answers to an end user. The question that arises here is: How and to which extent can these systems explain themselves? Can they reason logically? Can they reason causally? Or are they, as other scientists suspect, just stochastic parrots [334]?

Further tinnitus research using ML The tinnitus related outcomes of chapters 3.1 and 3.2 confirmed or extended the knowledge about tinnitus. Further studies could investigate to what extent temperature, humidity, or a confounder variable such as season have actual influence on tinnitus and what **biological mechanisms** are behind it. Regarding sleep problems in women, the direction of the cause would be intriguing: does tinnitus lead to sleep problems or does fatigue lead to tinnitus?

Increasing data quality and multi-modality in EMA and MCS using clinical studies The anonymity and non-binding nature of EMA studies has the advantage of reaching many different individuals. The disadvantage, however, is that the data obtained cannot be linked to others where the identity of the patients is known, such as in clinical trials. So, it would be exciting to see if EMA data could be augmented with biological data such as blood counts or physical tests by the treating physician to get a holistic picture of a disease process. Also, the exponential user drop off could be reduced, because there would be more liability and more benefits for the patient.

Implementing AI in health Not excluding our studies, most ML papers are not implemented. On the one hand, this is due to the lack of scope: the papers simply do not aim to produce proof-of-concepts, but remain in theory. On the other hand, it is also due to the fact that the implementation of AI systems is very complex and requires additional domain experts from software development (backend, frontend, full-stack developers) and IT security to make the systems secure against malicious attacks. There are already frameworks that outline requirements for the system: They should be purposeful, effective, safe, secure, private, fair, equitable, transparent, explainable, accountable, and monitored [335; 336]. That in mind, another dimension adds complexity: **Ethics**. For instance, an informed consent. Who should be informed that AI is working in the background? Which stakeholders have an interest in knowing? Which rights of patients should be protected and how? These issues are partially touched upon in the literature [337]. What is missing are guidelines and lessons learned for the implementation of AI systems in German or European hospitals. Also, when it comes

to practical implementation, the usefulness of a prediction might be questioned. For instance, the **prediction horizon** should be discussed: Is it useful to predict something 10 years ahead? Or 1 year? Or 1 hour? It depends on the **clinical workflow** and whether such a time window finally improves patient treatment. However, most of the research papers do not take this clinical workflow into account. As workflows might differ between hospitals, the implemented models differ and thus their will performance. What ultimately matters, however, is the local validity of the models, so a test set must ultimately be out-of-distribution and from the local clinic in order to assess the implementability of a model. For high-stake decisions, there should be a **domain expert in the loop**. But, when and how is this expert included? Is there a feedback loop to the model that runs in production? What information should be displayed on a dashboard? And does this ultimately improve the clinical workflow, or does it not stress the treating physician because another duty is added?

These are some research directions that we consider promising and valuable for the digital medicine community. In the last and final section, we would like to conclude this thesis.

4.5 | Conclusion

This work is affiliated with the **intersection of several disciplines** and includes the concepts of EMA, mHealth, Mobile Crowdsensing (MCS), Supervised Machine Learning, Machine Learning Explainability, and works with data from the Tinnitus Domain, Psychology, and the coronavirus. EMA involves the repeated sampling of a user's current experience and in his or her natural environment in real time. mHealth refers to medical procedures and private and public healthcare interventions delivered on mobile devices. MCS means the measurement and collection of data through different types of measurement devices (e.g., smartphones) by a large number of users. Supervised Machine Learning is a non-parametric mapping of an input A to an output B with known labels of B . Machine Learning Explainability describes methods that enable humans to understand why a model makes certain predictions. And the domain expert is a person with high level of expertise in the area of interest.

Within these fields, we originally asked **how can we achieve XAI within the mHealth domain and if yes, can we beneficially use XAI methods?** I then asked, in general terms, (1) what challenges are specific to EMA and MCS, (2) which value ML adds to the analyses of this data, (3) and why XAI is needed in that context. The specific challenges that we

found in the mHealth context (power user, no equidistant measurements, concept drift, missing values, operating-system-related measuring inaccuracies, uncertain user identity, different user behavior) are results of experience over time, and iterative communication with different domain experts from backend, app developers, application domain and data scientists. ML is a great tool among others and has the power to confirm or broaden domain knowledge, increase workflow efficiency, and improve medical treatment quality. However, its potential to add value depends on the use case and sometimes, statistics or even simpler heuristics outperform complex ML models. XAI is needed to improve model performance and robustness, and to build trust with stakeholders. The granularity, language, and complexity of a useful XAI method depends on the recipient of the explanation.

Altogether, one major lesson from this thesis is: The key of progression (gaining experience, finding and documenting pitfalls, meaningfully interpret data, and precisely draw conclusions) is an **open communication** between the specialists involved in the research project. For open communication, in turn, a common language during these projects is required. This can finally lead to strong bridges between medicine and informatics, resulting in new research areas like **digital medicine**, where people stand on the same side of the challenges and working to address new research questions together.

Appendix

The appendix contains a publication list with all papers of the doctoral candidate, affidavits in both German and English, a curriculum vitae, and contribution statements for papers, figures, tables, and a confirmation of legal second publication rights.

5.1 | Publication list

This is a list of publications of the doctoral candidate as per September, 8th, 2023. A current list of publications can also be found on [Google Scholar](#).

- [1] **Allgaier, Johannes**; Schlee, Winfried; Langguth, Berthold; Probst, Thomas; Pryss, Rüdiger. *Predicting the gender of individuals with tinnitus based on daily life data of the TrackYourTinnitus mHealth platform*. Scientific Reports, 11(1), 18375, 2021.
- [2] **Allgaier, Johannes**; Neff, Patrick; Schlee, Winfried; Schoisswohl, Stefan; Pryss, Rüdiger. *Deep learning end-to-end approach for the prediction of tinnitus based on EEG data*. 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 816–819, IEEE, 2021.
- [3] **Allgaier, Johannes**; Schlee, Winfried; Probst, Thomas; Pryss, Rüdiger. *Prediction of tinnitus perception based on daily life mhealth data using country origin and season*. Journal of Clinical Medicine, 11(15), 4270, 2022.
- [4] **Allgaier, Johannes**; Mulansky, Lena; Draelos, Rachel Lea; Pryss, Rüdiger. *How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare*. Artificial Intelligence in Medicine, 143, 102616, 2023.

- [5] **Allgaier, Johannes.** *Machine learning under concept drift for industrial data using Python*. Master Thesis, Institute of Databases and Information Systems, 2019.
- [6] Beierle, Felix; Schobel, Johannes; Vogel, Carsten; **Allgaier, Johannes**; Mulansky, Lena; Haug, Fabian; Haug, Julian; Schlee, Winfried; Holfelder, Marc; Stach, Michael; others. *Corona health—A study-and sensor-based mobile app platform exploring aspects of the COVID-19 pandemic*. *International Journal of Environmental Research and Public Health*, 18(14), 7395, 2021.
- [7] Beierle, Felix; **Allgaier, Johannes**; Stupp, Carolin; Keil, Thomas; Schlee, Winfried; Schobel, Johannes; Vogel, Carsten; Haug, Fabian; Haug, Julian; Holfelder, Marc; others. *Self-Assessment of Having COVID-19 With the Corona Check Mhealth App*. *IEEE Journal of Biomedical and Health Informatics*, 2023.
- [8] Kammerer, Klaus; Hoppenstedt, Burkhard; Pryss, Rüdiger; Stökler, Steffen; **Allgaier, Johannes**; Reichert, Manfred. *Anomaly detections for manufacturing systems based on sensor data—insights into two challenging real-world production settings*. *Sensors*, 19(24), 5370, 2019.
- [9] Schlee, Winfried; Schoisswohl, Stefan; Staudinger, Susanne; Schiller, Axel; Lehner, Astrid; Langguth, Berthold; Schecklmann, Martin; Simoes, Jorge; Neff, Patrick; Marcrum, Steven C; others. *Towards a unification of treatments and interventions for tinnitus patients: The EU research and innovation action UNITI*. *Progress in Brain Research*, 260, 441–451, 2021.
- [10] Schlee, Winfried; Langguth, Berthold; Pryss, Rüdiger; **Allgaier, Johannes**; Mulansky, Lena; Vogel, Carsten; Spiliopoulou, Myra; Schleicher, Miro; Unnikrishnan, Vishnu; Puga, Clara; others. *Using big data to develop a clinical decision support system for tinnitus treatment*. *The Behavioral Neuroscience of Tinnitus*, pp. 175–189, 2021.
- [11] Landauer, Jürgen; Hoppenstedt, Burkhard; **Allgaier, Johannes.** *Image segmentation to locate ancient Maya architectures using deep learning*. Publishers Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia, 7, 2022.
- [12] Fleischer, Anna; Heimeshoff, Larissa; **Allgaier, Johannes**; Jordan, Karin; Gelbrich, Götz; Pryss, Rüdiger; Schobel, Johannes; Einsele, Hermann; Kortuem, Martin; Maatouk, Imad; others. *Is PFS the right endpoint to assess outcome of maintenance studies in multiple myeloma? Results of a patient survey highlight quality-of-life as an equally important outcome measure*. *Blood*, 138, 836, 2021.

- [13] Fleischer, Anna; Zapf, Larissa; **Allgaier, Johannes**; Jordan, Karin; Gelbrich, Götz; Pryss, Rüdiger; Schobel, Johannes; Bittrich, Max; Einsele, Hermann; Kortüm, Martin; others. *A patient survey indicates quality of life and progression-free survival as equally important outcome measures in multiple myeloma clinical trials*. 2023.
- [14] Breitmayer, Marius; Stach, Michael; Kraft, Robin; **Allgaier, Johannes**; Reichert, Manfred; Schlee, Winfried; Probst, Thomas; Langguth, Berthold; Pryss, Rüdiger. *Predicting the presence of tinnitus using ecological momentary assessments*. *Scientific Reports*, 13(1), 8989, 2023.
- [15] IN REVIEW: **Allgaier, Johannes**; Pryss, Rüdiger. *7 observational mHealth studies and 10 years of experience: Can ignoring groups in Machine Learning pipelines lead to overestimation of model performance? Analyses of group-wise validation as well as baseline and concept-drift considerations*. *Nature Communications Medicine*, 2023.

5.2 | Affidavits

Affidavit I hereby confirm that my thesis entitled *Machine Learning Explainability on Multi-Modal Data using Ecological Momentary Assessments in the Medical Domain* is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis. Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Place, Date

Signature

Eidesstattliche Erklärung Hiermit erkläre ich an Eides statt, die Dissertation *Erklärbarkeit von maschinellem Lernen unter Verwendung multi-modaler Daten und Ecological Momentary Assessments im medizinischen Sektor* eigenständig, das heißt insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Place, Date

Signature

5.3 | Curriculum Vitae

5.4 | Contribution Statements

Statement of individual author contributions and of legal second publication rights

The following section provides one table for each manuscript indicating the contribution of each author with author initials, and responsibility decreasing from left to right.

Participated in	Author 1	Author 2	Author 3	Author 4	Author 5
Study Design	R. P.	J. A.	W. S.	T. P.	B. L.
Methods Development	R. P.	J. A.	W. S.	T. P.	B. L.
Data Collection	B. L.	R. P.	J. A.	W. S.	T. P.
Data Analysis and Interpretation	J. A.	R. P.	W. S.	T. P.	B. L.
Manuscript Writing	J. A.	R. P.	W. S.	T. P.	B. L.
Writing of Introduction	J. A.	R. P.	W. S.	T. P.	B. L.
Writing of Materials & Methods	J. A.	R. P.	W. S.	T. P.	B. L.
Writing of Discussion	J. A.	R. P.	W. S.	T. P.	B. L.
Writing of First Draft	J. A.	R. P.	W. S.	T. P.	B. L.

Table 5.1: Author contribution statement for manuscript 3.1.

Participated in	Author 1	Author 2	Author 3	Author 4
Study Design	R. P.	J. A.	W. S.	T. P.
Methods Development	R. P.	J. A.	W. S.	T. P.
Data Collection	B. L.	R. P.	J. A.	W. S.
Data Analysis and Interpretation	J. A.	R. P.	W. S.	T. P.
Manuscript Writing	J. A.	R. P.	W. S.	T. P.
Writing of Introduction	J. A.	R. P.	W. S.	T. P.
Writing of Materials & Methods	J. A.	R. P.	W. S.	T. P.
Writing of Discussion	J. A.	R. P.	W. S.	T. P.
Writing of First Draft	J. A.	R. P.	W. S.	T. P.

Table 5.2: Author contribution statement for manuscript 3.2.

Participated in	Author 1	Author 2	Author 3	Author 4
Study Design	R. P.	F. B.	C. C.	et. al.
Methods Development	F. B.	R. P.	J. A.	et. al.
Data Collection	R. P.	C. V.	J. S.	et. al.
Data Analysis and Interpretation	F. B.	J. A.	R. P.	et. al.
Manuscript Writing	F. B.	J. A.	R. P.	et. al.
Writing of Introduction	F. B.	J. A.	R. P.	et. al.
Writing of Materials & Methods	F. B.	J. A.	R. P.	et. al.
Writing of Discussion	F. B.	J. A.	R. P.	et. al.
Writing of First Draft	F. B.	J. A.	R. P.	et. al.

Table 5.3: Author contribution statement for manuscript 3.3.

Participated in	Author 1	Author 2	Author 3	Author 4
Study Design	J. A.	L. M.	R. D.	R. P.
Methods Development	J. A.	L. M.	R. D.	R. P.
Data Collection	J. A.	L. M.		
Data Analysis and Interpretation	J. A.	L. M.	R. D.	R. P.
Manuscript Writing	J. A.	L. M.	R. D.	R. P.
Writing of Introduction	J. A.	L. M.	R. D.	R. P.
Writing of Materials & Methods	J. A.	L. M.	R. D.	R. P.
Writing of Discussion	J. A.	L. M.	R. D.	R. P.
Writing of First Draft	J. A.	L. M.	R. D.	R. P.

Table 5.4: Author contribution statement for manuscript 3.4.

Participated in	Author 1	Author 2
Study Design	J. A.	R. P.
Methods Development	J. A.	R. P.
Data Collection	R. P.	J. A.
Data Analysis and Interpretation	J. A.	R. P.
Manuscript Writing	J. A.	R. P.
Writing of Introduction	J. A.	R. P.
Writing of Materials & Methods	J. A.	R. P.
Writing of Discussion	J. A.	R. P.
Writing of First Draft	J. A.	R. P.

Table 5.5: Author contribution statement for manuscript 3.5

If applicable, the doctoral researcher confirms that he has obtained permission from both the publishers (copyright) and the co-authors for legal second publication. The doctoral researcher and the primary supervisor confirm the correctness of the above mentioned assessment.

Würzburg, den 7. September 2023

Place, Date

Prof. Dr. Rüdiger Pryss

Place, Date

Johannes Allgaier

Statement of individual author contributions to figures/tables

The following section provides one table for each manuscript indicating the contribution of each author with author initials, and responsibility decreasing from left to right.

Figure / Table	Author 1	Author 2	Author 3	Author 4	Author 5
Table 3.1	J. A.	R. P.	W. S.	T. P.	B. L.
Figure 3.1	J. A.	R. P.	W. S.	T. P.	B. L.
Figure 3.2	J. A.	R. P.	W. S.	T. P.	B. L.
Figure 3.4	J. A.	R. P.	W. S.	T. P.	B. L.
Figure 3.5	J. A.	R. P.	W. S.	T. P.	B. L.
Figure 3.3	J. A.	R. P.	W. S.	T. P.	B. L.
Figure 3.6	J. A.	R. P.	W. S.	T. P.	B. L.
Figure 3.7	J. A.	R. P.	W. S.	T. P.	B. L.
Table 3.3	J. A.	R. P.	W. S.	T. P.	B. L.
Table 3.4	J. A.	R. P.	W. S.	T. P.	B. L.
Table 3.2	J. A.	R. P.	W. S.	T. P.	B. L.

Table 5.6: Contribution statements for manuscript 3.1 with author initials, and responsibility decreasing from left to right.

Figure / Table	Author 1	Author 2	Author 3	Author 4
Figure 3.12	J. A.	W. S.	T. P.	R. P.
Figure 3.9	J. A.	W. S.	T. P.	R. P.
Figure 3.11	J. A.	W. S.	T. P.	R. P.
Figure 3.10	J. A.	W. S.	T. P.	R. P.
Figure 3.13	J. A.	W. S.	T. P.	R. P.
Table 3.8	J. A.	W. S.	T. P.	R. P.
Table 3.10	J. A.	W. S.	T. P.	R. P.
Table 3.6	J. A.	W. S.	T. P.	R. P.
Table 3.5	J. A.	W. S.	T. P.	R. P.
Table 3.7	J. A.	W. S.	T. P.	R. P.
Table 3.9	J. A.	W. S.	T. P.	R. P.

Table 5.7: Contribution statements for manuscript 3.2 with author initials, and responsibility decreasing from left to right.

Figure / Table	Author 1	Author 2	Author 3	Author 4
Figure 3.17	F. B.	J. A.	R. P.	et. al.
Figure 3.16	F. B.	J. A.	R. P.	et. al.
Figure 3.15	F. B.	J. A.	R. P.	et. al.
Figure 3.19	J. A.	F. B.	R. P.	et. al.
Figure 3.18	J. A.	F. B.	R. P.	et. al.
Table 3.11	F. B.	J. A.	R. P.	et. al.
Table 3.13	J. A.	F. B.	R. P.	et. al.
Table 3.14	J. A.	F. B.	R. P.	et. al.
Table 3.12	J. A.	F. B.	R. P.	et. al.
Table 3.15	J. A.	F. B.	R. P.	et. al.
Table 3.16	J. A.	F. B.	R. P.	et. al.

Table 5.8: Contribution statements for manuscript 3.3 with author initials, and responsibility decreasing from left to right.

Figure / Table	Author 1	Author 2	Author 3	Author 4
Figure 3.28	J. A.	L. M.	R. D.	R. P.
Figure 3.29	J. A.	L. M.	R. D.	R. P.
Figure 3.26	J. A.	L. M.	R. D.	R. P.
Figure 3.25	J. A.	L. M.	R. D.	R. P.
Figure 3.27	J. A.	L. M.	R. D.	R. P.
Figure 3.22	J. A.	L. M.	R. D.	R. P.
Figure 3.20	J. A.	L. M.	R. D.	R. P.
Figure 3.23	J. A.	L. M.	R. D.	R. P.
Figure 3.24	J. A.	L. M.	R. D.	R. P.
Table 3.18	J. A.	L. M.	R. D.	R. P.
Table 3.19	J. A.	L. M.	R. D.	R. P.
Table 3.17	J. A.	L. M.	R. D.	R. P.

Table 5.9: Contribution statements for manuscript 3.4 with author initials, and responsibility decreasing from left to right.

Figure / Table	Author 1	Author 2
Figure 3.31	J. A.	R. P.
Figure 3.32	J. A.	R. P.
Figure 3.33	J. A.	R. P.
Figure 3.30	J. A.	R. P.
Figure 3.24	J. A.	R. P.
Figure 3.34	J. A.	R. P.
Table 3.22	J. A.	R. P.
Table 3.25	J. A.	R. P.
Table 3.20	J. A.	R. P.
Table 3.21	J. A.	R. P.
Table 3.23	J. A.	R. P.

Table 5.10: Contribution statements for manuscript 3.5 with author initials, and responsibility decreasing from left to right.

I, Johannes Allgaier, also confirm my primary supervisor’s acceptance.

Place, Date

Johannes Allgaier

References

- [1] J. Allgaier, L. Mulansky, R. L. Draelos, and R. Pryss, "How does the model make predictions? a systematic literature review on the explainability power of machine learning in healthcare," *Artificial Intelligence in Medicine*, vol. 143, p. 102616, 2023.
- [2] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment," *Annu. Rev. Clin. Psychol.*, vol. 4, pp. 1–32, 2008.
- [3] A. A. Stone and S. Shiffman, "Ecological momentary assessment (ema) in behavioral medicine." *Annals of behavioral medicine*, 1994.
- [4] R. Pryss, M. Reichert, J. Herrmann, B. Langguth, and W. Schlee, "Mobile crowd sensing in clinical and psychological trials—a case study," in *2015 IEEE 28th international symposium on computer-based medical systems*. IEEE, 2015, pp. 23–24.
- [5] S. Shiffman, "Dynamic influences on smoking relapse process," *Journal of personality*, vol. 73, no. 6, pp. 1715–1748, 2005.
- [6] S. Shiffman, J. A. Paty, M. Gnys, J. A. Kassel, and M. Hickcox, "First lapses to smoking: within-subjects analysis of real-time reports." *Journal of consulting and clinical psychology*, vol. 64, no. 2, p. 366, 1996.
- [7] R. Pryss, W. Schlee, B. Langguth, and M. Reichert, "Mobile crowdsensing services for tinnitus assessment and patient feedback," in *2017 IEEE International Conference on AI & Mobile Services (AIMS)*. IEEE, 2017, pp. 22–29.
- [8] F. Beierle, J. Schobel, C. Vogel, J. Allgaier, L. Mulansky, F. Haug, J. Haug, W. Schlee, M. Holfelder, M. Stach *et al.*, "Corona health—a study-and sensor-based mobile app platform exploring aspects of the covid-19 pandemic," *International Journal of Environmental Research and Public Health*, vol. 18, no. 14, p. 7395, 2021.
- [9] S. Akter and P. Ray, "mhealth—an ultimate platform to serve the unserved," *Yearbook of medical informatics*, vol. 19, no. 01, pp. 94–100, 2010.
- [10] B. Martínez-Pérez, I. De La Torre-Díez, and M. López-Coronado, "Mobile health applications for the most prevalent conditions by the world health organization: review and analysis," *Journal of medical Internet research*, vol. 15, no. 6, p. e120, 2013.

- [11] D. M. El-Sherif and M. Abouzid, "Analysis of mhealth research: mapping the relationship between mobile apps technology and healthcare during covid-19 outbreak," *Globalization and Health*, vol. 18, no. 1, p. 67, 2022.
- [12] A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, and Y. Luo, "Multimodal machine learning in precision health: A scoping review," *npj Digital Medicine*, vol. 5, no. 1, p. 171, 2022.
- [13] P. Giordani, S. Perna, A. Bianchi, A. Pizzulli, S. Tripodi, and P. M. Matricardi, "A study of longitudinal mobile health data through fuzzy clustering methods for functional data: The case of allergic rhinoconjunctivitis in childhood," *Plos one*, vol. 15, no. 11, p. e0242197, 2020.
- [14] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: current state and future challenges," *IEEE communications Magazine*, vol. 49, no. 11, pp. 32–39, 2011.
- [15] R. Pryss, W. Schlee, B. Hoppenstedt, M. Reichert, M. Spiliopoulou, B. Langguth, M. Breitmayer, T. Probst *et al.*, "Applying machine learning to daily-life data from the trackyourtinnitus mobile health crowdsensing platform to predict the mobile operating system used with high accuracy: Longitudinal observational study," *Journal of Medical Internet Research*, vol. 22, no. 6, p. e15547, 2020.
- [16] D. A. Waterman, *A guide to expert systems*. Addison-Wesley Longman Publishing Co., Inc., 1985.
- [17] S. Russel, P. Norvig *et al.*, *Artificial intelligence: a modern approach*. Pearson Education Limited London, 2013, vol. 256.
- [18] S. B. Kotsiantis, I. Zaharakis, P. Pintelas *et al.*, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.
- [19] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2021.
- [20] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.
- [21] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.
- [22] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.
- [23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.
- [24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [25] C. Bellarosa and P. Y. Chen, "The effectiveness and practicality of occupational stress management interventions: A survey of subject matter expert opinions." *Journal of occupational health psychology*, vol. 2, no. 3, p. 247, 1997.

- [26] M. Allen, R. Leung, J. McGrenere, and B. Purves, "Involving domain experts in assistive technology research," *Universal Access in the Information Society*, vol. 7, pp. 145–154, 2008.
- [27] R. Kraft, M. Reichert, and R. Pryss, "Towards the interpretation of sound measurements from smart-phones collected with mobile crowdsensing in the healthcare domain: An experiment with android devices," *Sensors*, vol. 22, no. 1, p. 170, 2022.
- [28] K. Kammerer, B. Hoppenstedt, R. Pryss, S. Stökler, J. Allgaier, and M. Reichert, "Anomaly detections for manufacturing systems based on sensor data—insights into two challenging real-world production settings," *Sensors*, vol. 19, no. 24, p. 5370, 2019.
- [29] R. Pryss, W. Schlee, B. Hoppenstedt, M. Reichert, M. Spiliopoulou, B. Langguth, M. Breitmayer, T. Probst *et al.*, "Applying machine learning to daily-life data from the trackyourtinnitus mobile health crowdsensing platform to predict the mobile operating system used with high accuracy: Longitudinal observational study," *Journal of Medical Internet Research*, vol. 22, no. 6, p. e15547, 2020.
- [30] R. Kraft, W. Schlee, M. Stach, M. Reichert, B. Langguth, H. Baumeister, T. Probst, R. Hannemann, and R. Pryss, "Combining mobile crowdsensing and ecological momentary assessments in the healthcare domain," *Frontiers in neuroscience*, vol. 14, p. 164, 2020.
- [31] J. Allgaier, W. Schlee, T. Probst, and R. Pryss, "Prediction of tinnitus perception based on daily life mhealth data using country origin and season," *Journal of Clinical Medicine*, vol. 11, no. 15, p. 4270, 2022.
- [32] K. Nelson, G. Corbin, M. Anania, M. Kovacs, J. Tobias, and M. Blowers, "Evaluating model drift in machine learning algorithms," in *2015 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*. IEEE, 2015, pp. 1–8.
- [33] A. Tsymbal, "The problem of concept drift: definitions and related work," *Computer Science Department, Trinity College Dublin*, vol. 106, no. 2, p. 58, 2004.
- [34] A. C. Acock, "Working with missing values," *Journal of Marriage and family*, vol. 67, no. 4, pp. 1012–1028, 2005.
- [35] M. Hölzl, R. Neumeier, and G. Ostermayer, "Analysis of compass sensor accuracy on several mobile devices in an industrial environment," in *Computer Aided Systems Theory-EUROCAST 2013: 14th International Conference, Las Palmas de Gran Canaria, Spain, February 10-15, 2013. Revised Selected Papers, Part II 14*. Springer, 2013, pp. 381–389.
- [36] B. Bent, B. A. Goldstein, W. A. Kibbe, and J. P. Dunn, "Investigating sources of inaccuracy in wearable optical heart rate sensors," *NPJ digital medicine*, vol. 3, no. 1, p. 18, 2020.
- [37] F. Lieder, T. L. Griffiths, Q. J. M. Huys, and N. D. Goodman, "The anchoring bias reflects rational use of cognitive resources," *Psychonomic bulletin & review*, vol. 25, pp. 322–349, 2018.
- [38] E. Humer, T. Keil, C. Stupp, W. Schlee, M. Wildner, P. Heuschmann, M. Winter, T. Probst, R. Pryss *et al.*, "Associations of country-specific and sociodemographic factors with self-reported covid-19-related symptoms: Multivariable analysis of data from the coronacheck mobile health platform," *JMIR Public Health and Surveillance*, vol. 9, no. 1, p. e40958, 2023.

- [39] T. Glasmachers, "Limits of end-to-end learning," in *Asian conference on machine learning*. PMLR, 2017, pp. 17–32.
- [40] N. Amangeldiuly, D. Karlov, and M. V. Fedorov, "Baseline model for predicting protein–ligand unbinding kinetics through machine learning," *Journal of Chemical Information and Modeling*, vol. 60, no. 12, pp. 5946–5956, 2020.
- [41] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, 2019.
- [42] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in medicine*, vol. 23, no. 1, pp. 89–109, 2001.
- [43] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *Radiographics*, vol. 37, no. 2, pp. 505–515, 2017.
- [44] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Annals of translational medicine*, vol. 4, no. 11, 2016.
- [45] L. Breiman, *Classification and regression trees*. Routledge, 2017.
- [46] —, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.
- [47] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.
- [48] N. Altman and M. Krzywinski, "Points of significance: Association, correlation and causation." *Nature methods*, vol. 12, no. 10, 2015.
- [49] J. G. Richens, C. M. Lee, and S. Johri, "Improving the accuracy of medical diagnosis with causal machine learning," *Nature communications*, vol. 11, no. 1, p. 3923, 2020.
- [50] J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, and R. Silva, "Causal machine learning: A survey and open problems," *arXiv preprint arXiv:2206.15475*, 2022.
- [51] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.
- [52] J. Allgaier, W. Schlee, B. Langguth, T. Probst, and R. Pryss, "Predicting the gender of individuals with tinnitus based on daily life data of the trackyourtinnitus mhealth platform," *Scientific Reports*, vol. 11, no. 1, p. 18375, 2021.
- [53] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *Ieee Access*, vol. 8, pp. 42 200–42 216, 2020.
- [54] C. Schaffer, "Selecting a classification method by cross-validation," *Machine learning*, vol. 13, pp. 135–143, 1993.
- [55] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

- [56] D. Chakraborty, I. Awolusi, and L. Gutierrez, "An explainable machine learning model to predict and elucidate the compressive behavior of high-performance concrete," *Results in Engineering*, vol. 11, p. 100245, 2021.
- [57] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai," *International Journal of Human-Computer Studies*, vol. 146, p. 102551, 2021.
- [58] E. U. Commission *et al.*, "Regulatory framework proposal on artificial intelligence," 2021.
- [59] I. A. of Diabetes and P. S. G. C. Panel, "International association of diabetes and pregnancy study groups recommendations on the diagnosis and classification of hyperglycemia in pregnancy," *Diabetes care*, vol. 33, no. 3, pp. 676–682, 2010.
- [60] A. Y. Ng *et al.*, "Preventing "overfitting" of cross-validation data," in *ICML*, vol. 97. Citeseer, 1997, pp. 245–253.
- [61] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence," *Science robotics*, vol. 4, no. 37, p. eaay7120, 2019.
- [62] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.
- [63] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 2018, pp. 0210–0215.
- [64] J. F. Weaver, "The federal government and trustworthy ai," *The Journal of Robotics, Artificial Intelligence & Law*, vol. 4.
- [65] W. House, "Blueprint for an ai bill of rights—making automated systems work for the american people," 2022.
- [66] OECD, "The organization for economic cooperation and development: Ai principles," 2023.
- [67] Food, D. Administration *et al.*, "Proposed regulatory framework for modifications to artificial intelligence/machine learning (ai/ml)-based software as a medical device (samd)," 2019.
- [68] Food and U. Drug Administration of the United States, "Artificial intelligence and machine learning in software as a medical device," *FDA*, 2021.
- [69] U. FDA, "International medical device regulators forum (imdrf)," 2019.
- [70] C. Sorenson and M. Drummond, "Improving medical device regulation: the united states and europe in perspective," *The Milbank Quarterly*, vol. 92, no. 1, pp. 114–150, 2014.
- [71] D. B. Kramer, S. Xu, and A. S. Kesselheim, "How does medical device regulation perform in the united states and the european union? a systematic review," 2012.
- [72] D. Adrian and A. El Refaie, "The epidemiology of tinnitus," in *The handbook of tinnitus*. Singular, 2000, pp. 1–23.

- [73] D. De Ridder, A. B. Elgoyhen, R. Romo, and B. Langguth, "Phantom percepts: tinnitus and pain as persisting aversive memory networks," *Proceedings of the National Academy of Sciences*, vol. 108, no. 20, pp. 8075–8080, 2011.
- [74] R. J. Budd and R. Pugh, "Tinnitus coping style and its relationship to tinnitus severity and emotional distress," *Journal of psychosomatic research*, vol. 41, no. 4, pp. 327–335, 1996.
- [75] Y. H. Kim, "Seasonal affective disorder in patients with chronic tinnitus," *The Laryngoscope*, vol. 126, no. 2, pp. 447–451, 2016.
- [76] Y. Alimohamadi, M. Sepandi, M. Taghdir, and H. Hosamirudsari, "Determine the most common clinical symptoms in covid-19 patients: a systematic review and meta-analysis," *Journal of preventive medicine and hygiene*, vol. 61, no. 3, p. E304, 2020.
- [77] C. Vogel, R. Pryss, J. Schobel, W. Schlee, and F. Beierle, "Developing apps for researching the covid-19 pandemic with the trackyourhealth platform," in *2021 IEEE/ACM 8th International Conference on Mobile Software Engineering and Systems (MobileSoft)*. IEEE, 2021, pp. 65–68.
- [78] J. Schobel, "A model-driven framework for enabling flexible and robust mobile data collection applications," Ph.D. dissertation, Ulm University, 2018.
- [79] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.
- [80] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, 1959.
- [81] E. Briscoe and J. Feldman, "Conceptual complexity and the bias/variance tradeoff," *Cognition*, vol. 118, no. 1, pp. 2–16, 2011.
- [82] Y. Dar, V. Muthukumar, and R. G. Baraniuk, "A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning," *arXiv preprint arXiv:2109.02355*, 2021.
- [83] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
- [84] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. Springer, 2009.
- [85] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in neural information processing systems*, vol. 14, 2002, pp. 841–848.
- [86] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 2, 1995, pp. 1137–1145.
- [87] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *Journal of Machine Learning Research*, vol. 5, pp. 1089–1105, 2004.

- [88] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
- [89] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.
- [90] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," *Advances in neural information processing systems*, vol. 28, 2015.
- [91] A. Paleyes, R.-G. Urma, and N. D. Lawrence, "Challenges in deploying machine learning: a survey of case studies," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–29, 2022.
- [92] C. Shearer, "The crisp-dm model: the new blueprint for data mining," *Journal of data warehousing*, vol. 5, no. 4, pp. 13–22, 2000.
- [93] A. A. Verma, J. Murray, R. Greiner, J. P. Cohen, K. G. Shojania, M. Ghassemi, S. E. Straus, C. Pou-Prom, and M. Mamdani, "Implementing machine learning in medicine," *Cmaj*, vol. 193, no. 34, pp. E1351–E1357, 2021.
- [94] T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)?—arguments against avoiding rmse in the literature," *Geoscientific model development*, vol. 7, no. 3, pp. 1247–1250, 2014.
- [95] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, "Assessing the performance of prediction models: a framework for some traditional and novel measures," *Epidemiology (Cambridge, Mass.)*, vol. 21, no. 1, p. 128, 2010.
- [96] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Machine learning*, vol. 23, pp. 69–101, 1996.
- [97] A. Wong, E. Otlis, J. P. Donnelly, A. Krumm, J. McCullough, O. DeTroyer-Cooley, J. Pestrue, M. Phillips, J. Konye, C. Penzoza *et al.*, "External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients," *JAMA Internal Medicine*, vol. 181, no. 8, pp. 1065–1070, 2021.
- [98] P. G. Lyons, M. R. Hofford, S. C. Yu, A. P. Michelson, P. R. O. Payne, C. L. Hough, and K. Singh, "Factors Associated With Variability in the Performance of a Proprietary Sepsis Prediction Model Across 9 Networked Hospitals in the US," *JAMA Internal Medicine*, 04 2023. [Online]. Available: <https://doi.org/10.1001/jamainternmed.2022.7182>
- [99] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.
- [100] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning—a brief history, state-of-the-art and challenges," in *ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, September 14–18, 2020, Proceedings*. Springer, 2021, pp. 417–431.
- [101] A. Altmann, L. Tološi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.

- [102] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," *Advances in neural information processing systems*, vol. 26, pp. 431–439, 2013.
- [103] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.
- [104] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International conference on machine learning*. PMLR, 2017, pp. 3145–3153.
- [105] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.
- [106] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," *Advances in neural information processing systems*, vol. 29, 2016.
- [107] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu *et al.*, "Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy," *Ophthalmology*, vol. 126, no. 4, pp. 552–564, 2019.
- [108] M. T. Ribeiro, S. Singh, and C. Guestrin, "'why should i trust you?' explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
- [109] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
- [110] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1094–1111, 2018.
- [111] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1885–1894.
- [112] N. Frosst and G. Hinton, "Distilling a neural network into a soft decision tree," *arXiv e-prints*, pp. arXiv-1711, 2017.
- [113] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv e-prints*, pp. arXiv-1706, 2017.
- [114] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*. PMLR, 2018, pp. 2668–2677.
- [115] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

- [116] C.-K. Yeh, J. Kim, I. E.-H. Yen, and P. K. Ravikumar, "Representer point selection for explaining deep neural networks," *Advances in neural information processing systems*, vol. 31, 2018.
- [117] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [118] Y. Du, J. Leung, and Y. Shi, "Perturbationrank: A non-monotone ranking algorithm," 2008.
- [119] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [120] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.
- [121] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model," *Nature methods*, vol. 12, no. 10, pp. 931–934, 2015.
- [122] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv e-prints*, pp. arXiv-1702, 2017.
- [123] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Gradient-based attribution methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 169–191.
- [124] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1-22. Prague, 2004, pp. 1–2.
- [125] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Muller, and W. Samek, "Analyzing classifiers: Fisher vectors and deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2912–2920.
- [126] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.
- [127] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.
- [128] W. Nie, Y. Zhang, and A. Patel, "A theoretical explanation for perplexing behaviors of backpropagation-based visualizations," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3809–3818.
- [129] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.
- [130] Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell, "Why should you trust my explanation?" understanding uncertainty in lime explanations," *arXiv e-prints*, pp. arXiv-1904, 2019.
- [131] M. R. Zafar and N. Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 525–541, 2021.

- [132] Z. Zhou, G. Hooker, and F. Wang, "S-lime: Stabilized-lime for model explanation," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 2429–2438.
- [133] L. Shapley, "Notes on the n-person game—ii: The value of an n-person game, the rand corporation, the rand corporation," *Research Memorandum*, vol. 670, 1951.
- [134] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv preprint arXiv:1605.01713*, 2016.
- [135] O. Irsoy, O. T. Yildiz, and E. Alpaydin, "Soft decision trees," in *International Conference on Pattern Recognition*, 2012.
- [136] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3319–3328.
- [137] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv e-prints*, pp. arXiv–1412, 2014.
- [138] N. Kiang, E. Moxon, and R. Levine, "Auditory-nerve activity in cats with normal and abnormal cochleas," *Sensorineural hearing loss*, pp. 241–273, 1970.
- [139] A. Davis and E. A. Rafea, "Epidemiology of tinnitus," *Tinnitus handbook*, vol. 1, p. 23, 2000.
- [140] B. Langguth, "A review of tinnitus symptoms beyond 'ringing in the ears': a call to action," *Current medical research and opinion*, vol. 27, no. 8, pp. 1635–1643, 2011.
- [141] J. B. Halford and S. D. Anderson, "Anxiety and depression in tinnitus sufferers," *Journal of psychosomatic research*, vol. 35, no. 4-5, pp. 383–390, 1991.
- [142] M. Mehdi, A. Dode, R. Pryss, W. Schlee, M. Reichert, and F. J. Hauck, "Contemporary and systematic review of smartphone apps for tinnitus management and treatment," 2020.
- [143] C. R. Cederroth, S. Gallus, D. A. Hall, T. Kleinjung, B. Langguth, A. Maruotti, M. Meyer, A. Norena, T. Probst, R. Pryss *et al.*, "Towards an understanding of tinnitus heterogeneity," *Frontiers in aging neuroscience*, vol. 11, p. 53, 2019.
- [144] C. R. Cederroth, U. Albrecht, J. Bass, S. A. Brown, J. Dyhrfeld-Johnsen, F. Gachon, C. B. Green, M. H. Hastings, C. Helfrich-Förster, J. B. Hogenesch *et al.*, "Medicine in the fourth dimension," *Cell metabolism*, vol. 30, no. 2, pp. 238–250, 2019.
- [145] R. Tyler, C. Coelho, P. Tao, H. Ji, W. Noble, A. Gehringer, and S. Gogel, "Identifying tinnitus subgroups with cluster analysis," *American journal of audiology*, vol. 17, no. 2, pp. 176–184, 2008.
- [146] B. Langguth, M. Landgrebe, W. Schlee, M. Schecklmann, V. Vielsmeier, T. Steffens, S. Staudinger, H. Frick, and U. Frick, "Different patterns of hearing loss among tinnitus patients: a latent class analysis of a large sample," *Frontiers in neurology*, vol. 8, p. 46, 2017.
- [147] M. Schecklmann, A. Lehner, T. B. Poepl, P. M. Kreuzer, G. Hajak, M. Landgrebe, and B. Langguth, "Cluster analysis for identifying sub-types of tinnitus: a positron emission tomography and voxel-based morphometry study," *Brain research*, vol. 1485, pp. 3–9, 2012.

- [148] R. Tyler, H. Ji, A. Perreau, S. Witt, W. Noble, and C. Coelho, "Development and validation of the tinnitus primary function questionnaire," *American Journal of Audiology*, vol. 23, no. 3, pp. 260–272, 2014.
- [149] C. R. Cederroth and W. Schlee, "Sex and gender differences in tinnitus," *Frontiers in Neuroscience*, vol. 16, p. 59, 2022.
- [150] U. Niemann, B. Boecking, P. Brueggemann, B. Mazurek, and M. Spiliopoulou, "Gender-specific differences in patients with chronic tinnitus—baseline characteristics and treatment effects," *Frontiers in Neuroscience*, vol. 14, p. 487, 2020.
- [151] A. Van der Wal, T. Luyten, E. Cardon, L. Jacquemin, O. M. Vanderveken, V. Topsakal, P. Van de Heyning, W. De Hertogh, N. Van Looveren, V. Van Rompaey *et al.*, "Sex differences in the response to different tinnitus treatment," *Frontiers in Neuroscience*, vol. 14, p. 422, 2020.
- [152] T. S. Han, J.-E. Jeong, S.-N. Park, and J. J. Kim, "Gender differences affecting psychiatric distress and tinnitus severity," *Clinical Psychopharmacology and Neuroscience*, vol. 17, no. 1, p. 113, 2019.
- [153] J. van Os, S. Verhagen, A. Marsman, F. Peeters, M. Bak, M. Marcelis, M. Drukker, U. Reininghaus, N. Jacobs, T. Lataster *et al.*, "The experience sampling method as an mhealth tool to support self-monitoring, self-insight, and personalized health care in clinical practice," *Depression and anxiety*, vol. 34, no. 6, pp. 481–493, 2017.
- [154] J. Torous, R. Friedman, and M. Keshavan, "Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions," *JMIR mHealth and uHealth*, vol. 2, no. 1, p. e2, 2014.
- [155] S. P. Rowland, J. E. Fitzgerald, T. Holme, J. Powell, and A. McGregor, "What is the clinical value of mhealth for patients?" *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–6, 2020.
- [156] A. Seifert, M. Hofer, and M. Allemand, "Mobile data collection: Smart, but not (yet) smart enough," *Frontiers in Neuroscience*, vol. 12, p. 971, 2018.
- [157] R. Pryss *et al.*, "Exploring the time trend of stress levels while using the crowdsensing mobile health platform, trackyourstress, and the influence of perceived stress reactivity: ecological momentary assessment pilot study," *JMIR mHealth and uHealth*, vol. 7, no. 10, p. e13978, 2019.
- [158] R. Pryss, "Mobile crowdsensing in healthcare scenarios: taxonomy, conceptual pillars, smart mobile crowdsensing services," in *Digital Phenotyping and Mobile Sensing*. Springer, 2019, pp. 221–234.
- [159] W. Schlee, R. C. Pryss, T. Probst, J. Schobel, A. Bachmeier, M. Reichert, and B. Langguth, "Measuring the moment-to-moment variability of tinnitus: the trackyourtinnitus smart phone app," *Frontiers in aging neuroscience*, vol. 8, p. 294, 2016.
- [160] T. Probst, R. Pryss, B. Langguth, and W. Schlee, "Emotional states as mediators between tinnitus loudness and tinnitus distress in daily life: Results from the "trackyourtinnitus" application," *Scientific reports*, vol. 6, no. 1, pp. 1–8, 2016.
- [161] W. Schlee, R. Kraft, J. Schobel, B. Langguth, T. Probst, P. Neff, M. Reichert, and R. Pryss, "Momentary assessment of tinnitus—how smart mobile applications advance our understanding of tinnitus," in *Digital Phenotyping and Mobile Sensing*. Springer, 2019, pp. 209–220.

- [162] F. Beierle *et al.*, "What data are smartphone users willing to share with researchers?" *Journal of ambient intelligence and humanized computing*, pp. 1–13, 2019.
- [163] R. Kraft, M. Stach, M. Reichert, W. Schlee, T. Probst, B. Langguth, M. Schickler, H. Baumeister, and R. Pryss, "Comprehensive insights into the trackyourtinnitus database," 2020.
- [164] M. Sereda, S. Smith, K. Newton, and D. Stockdale, "Mobile apps for management of tinnitus: users' survey, quality assessment, and content analysis," *JMIR mHealth and uHealth*, vol. 7, no. 1, p. e10353, 2019.
- [165] M. Mehdi, C. Riha, P. Neff, A. Dode, R. Pryss, W. Schlee, M. Reichert, and F. J. Hauck, "Smartphone apps in the context of tinnitus: Systematic review," *Sensors*, vol. 20, no. 6, p. 1725, 2020.
- [166] Y. K. Cheung, P.-Y. S. Hsueh, M. Qian, S. Yoon, L. Meli, K. M. Diaz, J. E. Schwartz, I. M. Kronish, and K. W. Davidson, "Are nomothetic or ideographic approaches superior in predicting daily exercise behaviors? analyzing n-of-1 mhealth data," *Methods of information in medicine*, vol. 56, no. 6, p. 452, 2017.
- [167] V. Unnikrishnan, Y. Shah, M. Schleicher, M. Strandzheva, P. Dimitrov, D. Velikova, R. Pryss, J. Schobel, W. Schlee, and M. Spiliopoulou, "Predicting the health condition of mhealth app users with large differences in the number of recorded observations-where to learn from?" in *International Conference on Discovery Science*. Springer, 2020, pp. 659–673.
- [168] A. Aguilera, C. A. Figueroa, R. Hernandez-Ramos, U. Sarkar, A. Cembali, L. Gomez-Pathak, J. Miramontes, E. Yom-Tov, B. Chakraborty, X. Yan *et al.*, "mhealth app using machine learning to increase physical activity in diabetes and depression: clinical trial protocol for the diamante study," *BMJ open*, vol. 10, no. 8, p. e034723, 2020.
- [169] A. B. Said, A. Mohamed, T. Elfouly, K. Abualsaud, and K. Harras, "Deep learning and low rank dictionary model for mhealth data classification," in *2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC)*. IEEE, 2018, pp. 358–363.
- [170] K. N. Qureshi, S. Din, G. Jeon, and F. Piccialli, "An accurate and dynamic predictive model for a smart m-health system using machine learning," *Information Sciences*, vol. 538, pp. 486–502, 2020.
- [171] V. Unnikrishnan, C. Beyer, P. Matuszyk, U. Niemann, R. Pryss, W. Schlee, E. Ntoutsis, and M. Spiliopoulou, "Entity-level stream classification: exploiting entity similarity to label the future observations referring to an entity," *International Journal of Data Science and Analytics*, vol. 9, no. 1, pp. 1–15, 2020.
- [172] R. Pryss, T. Probst, W. Schlee, J. Schobel, B. Langguth, P. Neff, M. Spiliopoulou, and M. Reichert, "Prospective crowdsensing versus retrospective ratings of tinnitus variability and tinnitus–stress associations based on the trackyourtinnitus mobile platform," *International Journal of Data Science and Analytics*, vol. 8, no. 4, pp. 327–338, 2019.
- [173] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

- [174] A. R. Fekr, M. Janidarmian, K. Radecka, and Z. Zilic, "Respiration disorders classification with informative features for m-health applications," *IEEE journal of biomedical and health informatics*, vol. 20, no. 3, pp. 733–747, 2015.
- [175] F. Khatun, A. E. Heywood, S. M. A. Hanifi, M. S. Rahman, P. K. Ray, S.-T. Liaw, and A. Bhuiya, "Gender differentials in readiness and use of mhealth services in a rural area of bangladesh," *BMC health services research*, vol. 17, no. 1, p. 573, 2017.
- [176] D. Cirillo, S. Catuara-Solarz, C. Morey, E. Guney, L. Subirats, S. Mellino, A. Gigante, A. Valencia, M. J. Rementeria, A. S. Chadha *et al.*, "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–11, 2020.
- [177] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
- [178] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.
- [179] M. Pal, "Random forest classifier for remote sensing classification," *International journal of remote sensing*, vol. 26, no. 1, pp. 217–222, 2005.
- [180] D. Lavanya and K. U. Rani, "Performance evaluation of decision tree classifiers on medical datasets," *International Journal of Computer Applications*, vol. 26, no. 4, pp. 1–4, 2011.
- [181] S. Siu, G. Gibson, and C. Cowan, "Decision feedback equalisation using neural network structures and performance comparison with standard architecture," *IEE Proceedings I-Communications, Speech and Vision*, vol. 137, no. 4, pp. 221–225, 1990.
- [182] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler *et al.*, "Api design for machine learning software: experiences from the scikit-learn project," *arXiv preprint arXiv:1309.0238*, 2013.
- [183] P. Lameski, E. Zdravevski, R. Mingov, and A. Kulakov, "Svm parameter tuning with grid search and its impact on reduction of model over-fitting," in *Rough sets, fuzzy sets, data mining, and granular computing*. Springer, 2015, pp. 464–474.
- [184] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.
- [185] —, "Asymptotics for and against cross-validation," *Biometrika*, pp. 29–35, 1977.
- [186] F. Mosteller, J. W. Tukey *et al.*, *Data analysis and regression: a second course in statistics*, 1977.
- [187] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 3, pp. 131–156, 1997.
- [188] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 5–32, 1996.

- [189] R. S. Tyler and L. J. Baker, "Difficulties experienced by tinnitus sufferers," *Journal of Speech and Hearing disorders*, vol. 48, no. 2, pp. 150–154, 1983.
- [190] S. Vanneste, K. Joos, and D. De Ridder, "Prefrontal cortex based sex differences in tinnitus perception: same tinnitus intensity, same tinnitus distress, different mood," *PLoS One*, vol. 7, no. 2, p. e31182, 2012.
- [191] L. Basso, B. Boecking, P. Brueggemann, N. L. Pedersen, B. Canlon, C. R. Cederroth, and B. Mazurek, "Gender-specific risk factors and comorbidities of bothersome tinnitus," *Frontiers in neuroscience*, vol. 14, p. 706, 2020.
- [192] A. Fioretti, E. Natalini, D. Riedl, R. Moschen, and A. Eibenstein, "Gender comparison of psychological comorbidities in tinnitus patients—results of a cross-sectional study," *Frontiers in Neuroscience*, vol. 14, p. 704, 2020.
- [193] C. Seydel, H. Haupt, H. Olze, A. J. Szczepek, and B. Mazurek, "Gender and chronic tinnitus: differences in tinnitus-related distress depend on age and duration of tinnitus," *Ear and hearing*, vol. 34, no. 5, pp. 661–672, 2013.
- [194] K. Richter, M. Zimni, I. Tomova, L. Retzer, J. Höfig, S. Kellner, C. Fries, K. Bernstein, W. Hitzl, T. Hillemacher *et al.*, "Insomnia associated with tinnitus and gender differences," *International Journal of Environmental Research and Public Health*, vol. 18, no. 6, p. 3209, 2021.
- [195] M. Ciman, K. Wac *et al.*, "Smartphones as sleep duration sensors: validation of the isensesleep algorithm," *JMIR mHealth and uHealth*, vol. 7, no. 5, p. e11930, 2019.
- [196] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.
- [197] S. v. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of statistical software*, pp. 1–68, 2010.
- [198] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.
- [199] S. Kokoska and D. Zwillinger, *CRC standard probability and statistics tables and formulae*. Crc Press, 2000.
- [200] R. F. Tate, "Correlation between a discrete and a continuous variable. point-biserial correlation," *The Annals of mathematical statistics*, vol. 25, no. 3, pp. 603–607, 1954.
- [201] W. Bergsma, "A bias-correction for cramér's χ^2 and tschuprow's t ," *Journal of the Korean Statistical Society*, vol. 42, no. 3, pp. 323–328, 2013.
- [202] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.
- [203] G. E. Hinton, "Connectionist learning procedures. artificial intelligence, 40 1-3: 185 234, 1989. reprinted in j. carbonell, editor," *Machine Learning: Paradigms and Methods*, MIT Press, 1990.

- [204] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees. belmont, ca: Wadsworth," *International Group*, vol. 432, pp. 151–166, 1984.
- [205] B. Langguth, P. M. Kreuzer, T. Kleinjung, and D. De Ridder, "Tinnitus: causes and clinical management," *The Lancet Neurology*, vol. 12, no. 9, pp. 920–930, 2013.
- [206] K. Izuhara, K. Wada, K. Nakamura, Y. Tamai, M. Tsuji, Y. Ito, and C. Nagata, "Association between tinnitus and sleep disorders in the general japanese population," *Annals of Otolaryngology, Rhinology & Laryngology*, vol. 122, no. 11, pp. 701–706, 2013.
- [207] L. McKENNA, R. S. HALLAM, and R. HINCHCLIFFE, "The prevalence of psychological disturbance in neuro-otology outpatients," *Clinical Otolaryngology & Allied Sciences*, vol. 16, no. 5, pp. 452–456, 1991.
- [208] D. T. Plante and D. G. Ingram, "Seasonal trends in tinnitus symptomatology: evidence from internet search engine query data," *European Archives of Oto-Rhino-Laryngology*, vol. 272, no. 10, pp. 2807–2813, 2015.
- [209] A. C. Yang, N. E. Huang, C.-K. Peng, and S.-J. Tsai, "Do seasons have an influence on the incidence of depression? the use of an internet search engine query data as a proxy of human affect," *PloS one*, vol. 5, no. 10, p. e13728, 2010.
- [210] J. A. Hilger, "Autonomic dysfunction in the inner ear," *The Laryngoscope*, vol. 59, no. 1, pp. 1–11, 1949.
- [211] M. Atkinson, "Tinnitus aurium: some considerations concerning its origin and treatment," *Archives of otolaryngology*, vol. 45, no. 1, pp. 68–76, 1947.
- [212] A. L. Miller, "Epidemiology, etiology, and natural treatment of seasonal affective disorder." *Alternative medicine review*, vol. 10, no. 1, 2005.
- [213] S. H. Jain, B. W. Powers, J. B. Hawkins, and J. S. Brownstein, "The digital phenotype," *Nature biotechnology*, vol. 33, no. 5, pp. 462–463, 2015.
- [214] V. Unnikrishnan, M. Schleicher, Y. Shah, N. Jamaludeen, R. Pryss, J. Schobel, R. Kraft, W. Schlee, and M. Spiliopoulou, "The effect of non-personalised tips on the continued use of self-monitoring mhealth applications," *Brain Sciences*, vol. 10, no. 12, p. 924, 2020.
- [215] Z. Jafari, B. E. Kolb, and M. H. Mohajerani, "Age-related hearing loss and tinnitus, dementia risk, and auditory amplification outcomes," *Ageing research reviews*, vol. 56, p. 100963, 2019.
- [216] "Expert System for COVID-19 Diagnosis, author = Salman, F. M. and Abu-Naser, S. S., journal = International Journal of Academic Information Systems Research (IJAISR), volume = 4, number = 3, pages = 13, year = 2020."
- [217] H. R. Almadhoun and S. S. Abu-Naser, "An expert system for diagnosing coronavirus (covid-19) using s15," *International Journal of Academic Engineering Research (IJAER)*, vol. 4, no. 4, p. 9, 2020.
- [218] W. S. Erazo, G. P. Guerrero, C. C. Betancourt, and I. S. Salazar, "Chatbot implementation to collect data on possible covid-19 cases and release the pressure on the primary health care system," in *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Nov. 2020, pp. 0302–0307.

- [219] M. Khan, M. T. Mehran, Z. U. Haq, Z. Ullah, S. R. Naqvi, M. Ihsan, and H. Abbass, "Applications of artificial intelligence in COVID-19 pandemic: A comprehensive review," *Expert Systems with Applications*, vol. 185, p. 115695, Dec. 2021.
- [220] A. M. Ismael and A. Şengür, "Deep learning approaches for COVID-19 detection based on chest X-ray images," *Expert Systems with Applications*, vol. 164, p. 114054, Feb. 2021.
- [221] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, M. B. I. Reaz, and M. T. Islam, "Can AI Help in Screening Viral and COVID-19 Pneumonia?" *IEEE Access*, vol. 8, pp. 132 665–132 676, 2020.
- [222] "Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet, author = Panwar, H. and Gupta, P. K. and Siddiqui, M. K. and Morales-Menendez, R. and Singh, V., journal = Chaos, Solitons & Fractals, volume = 138, pages = 109944, year = 2020, month = sep."
- [223] S. Tabik, A. Gómez-Ríos, J. L. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, E. Guirado, J.-L. Suárez, J. Luengo, M. Valero-González *et al.*, "Covidgr dataset and covid-sdnet methodology for predicting covid-19 based on chest x-ray images," *IEEE journal of biomedical and health informatics*, vol. 24, no. 12, pp. 3595–3605, 2020.
- [224] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. Jamalipour Soufi, "Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning," *Medical Image Analysis*, vol. 65, p. 101794, Oct. 2020.
- [225] "Choquet Integral and Coalition Game-Based Ensemble of Deep Learning Models for COVID-19 Screening From Chest X-Ray Images, author = Bhowal, P. and Sen, S. and Yoon, J. H. and Geem, Z. W. and Sarkar, R., journal = IEEE Journal of Biomedical and Health Informatics, volume = 25, number = 12, pages = 4328–4339, year = 2021, month = dec."
- [226] J. Laguarda, F. Hueto, and B. Subirana, "Covid-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 275–281, 2020.
- [227] K. S. Alqudaihi, N. Aslam, I. U. Khan, A. M. Almuhaideb, S. J. Alsunaidi, N. M. A. R. Ibrahim, F. A. Alhaidari, F. S. Shaikh, Y. M. Alsenbel, D. M. Alalharith, H. M. Alharthi, W. M. Alghamdi, and M. S. Alshahrani, "Cough sound detection and diagnosis using artificial intelligence techniques: Challenges and opportunities," *IEEE Access*, vol. 9, pp. 102 327–102 344, 2021.
- [228] M. Wiczorek, J. Siłka, and M. Woźniak, "Neural network powered covid-19 spread forecasting model," *Chaos, Solitons & Fractals*, vol. 140, p. 110203, Nov. 2020.
- [229] R. Sujath, J. M. Chatterjee, and A. E. Hassanien, "A machine learning forecasting model for covid-19 pandemic in india," *Stochastic Environmental Research and Risk Assessment*, vol. 34, no. 7, pp. 959–972, Jul. 2020.
- [230] S. Tuli, S. Tuli, R. Tuli, and S. S. Gill, "Predicting the growth and trend of covid-19 pandemic using machine learning and cloud computing," *Internet of Things*, vol. 11, p. 100222, Sep. 2020.
- [231] N. Ahmed, R. A. Michelin, W. Xue, S. Ruj, R. Malaney, S. S. Kanhere, A. Seneviratne, W. Hu, H. Janicke, and S. K. Jha, "A survey of covid-19 contact tracing apps," *IEEE Access*, vol. 8, pp. 134 577–134 601, 2020.

- [232] F. Beierle, U. Dhakal, C. Cohrdes, S. Eicher, and R. Pryss, "Public perception of the german covid-19 contact-tracing app corona-warn-app," in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, Jun. 2021, pp. 342–347.
- [233] K. Oyibo, K. S. Sahu, A. Oetomo, and P. P. Morita, "Factors influencing the adoption of contact tracing applications: Systematic review and recommendations," *Frontiers in Digital Health*, vol. 4, 2022.
- [234] S. Ussai, M. Pistis, E. Missoni, B. Formenti, B. Armocida, T. Pedrazzi, F. Castelli, L. Monasta, B. Lauria, and I. Mariani, "'immuni' and the national health system: Lessons learnt from the covid-19 digital contact tracing in italy," *International Journal of Environmental Research and Public Health*, vol. 19, no. 12, p. 7529, Jan. 2022.
- [235] C. Menni, A. M. Valdes, M. B. Freidin, C. H. Sudre, L. H. Nguyen, D. A. Drew, S. Ganesh, T. Varsavsky, M. J. Cardoso, J. S. El-Sayed Moustafa, A. Visconti, P. Hysi, R. C. E. Bowyer, M. Mangino, M. Falchi, J. Wolf, S. Ourselin, A. T. Chan, C. J. Steves, and T. D. Spector, "Real-time tracking of self-reported symptoms to predict potential covid-19," *Nature Medicine*, vol. 26, no. 7, pp. 1037–1040, Jul. 2020.
- [236] K. Klaser, E. J. Thompson, L. H. Nguyen, C. H. Sudre, M. Antonelli, B. Murray, L. S. Canas, E. Molteni, M. S. Graham, E. Kerfoot, L. Chen, J. Deng, A. May, C. Hu, A. Guest, S. Selvachandran, D. A. Drew, M. Modat, A. T. Chan, J. Wolf, T. D. Spector, A. Hammers, E. L. Duncan, S. Ourselin, and C. J. Steves, "Anxiety and depression symptoms after covid-19 infection: Results from the covid symptom study app," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 92, no. 12, pp. 1254–1258, Dec. 2021.
- [237] C. Menni, A. M. Valdes, L. Polidori, M. Antonelli, S. Penamakuri, A. Nogal, P. Louca, A. May, J. C. Figueiredo, C. Hu, E. Molteni, L. Canas, M. F. Österdahl, M. Modat, C. H. Sudre, B. Fox, A. Hammers, J. Wolf, J. Capdevila, A. T. Chan, S. P. David, C. J. Steves, S. Ourselin, and T. D. Spector, "Symptom prevalence, duration, and risk of hospital admission in individuals infected with sars-cov-2 during periods of omicron and delta variant dominance: A prospective observational study from the zoe covid study," *The Lancet*, vol. 399, no. 10335, pp. 1618–1624, Apr. 2022.
- [238] R. R. A. Hakim, E. Rusdi, and M. A. Setiawan, "Android based expert system application for diagnose covid-19 disease: Cases study of banyumas regency," *Journal of Intelligent Computing and Health Informatics (JICHI)*, vol. 1, no. 2, pp. 26–38, Sep. 2020.
- [239] H. R. Banjar, H. Alkhatabi, N. Alganmi, and G. I. Almouhana, "Prototype development of an expert system of computerized clinical guidelines for covid-19 diagnosis and management in saudi arabia," *International Journal of Environmental Research and Public Health*, vol. 17, no. 21, p. 8066, Jan. 2020.
- [240] M. R. Mufid, A. Basofi, S. Mawaddah, K. Khotimah, and N. Fuad, "Risk diagnosis and mitigation system of covid-19 using expert system and web scraping," in *2020 International Electronics Symposium (IES)*, Sep. 2020, pp. 577–583.
- [241] G. Battineni, N. Chintalapudi, and F. Amenta, "Ai chatbot design during an epidemic like the novel coronavirus," *Healthcare*, vol. 8, no. 2, p. 154, Jun. 2020.
- [242] H. Mukhtar, S. Rubaiee, M. Krichen, and R. Alroobaea, "An iot framework for screening of covid-19 using real-time data from wearable sensors," *International Journal of Environmental Research and Public Health*, vol. 18, no. 8, p. 4022, Jan. 2021.

- [243] M. S. Astriani, A. Kurniawan, and N. N. Qomariyah, "Covid-19 self-detection magic mirror with iot-based heart rate and temperature sensors," in *2021 2nd International Conference on Innovative and Creative Information Technology (ICITech)*, Sep. 2021, pp. 212–215.
- [244] J. Skibinska, R. Burget, A. Channa, N. Popescu, and Y. Koucheryavy, "Covid-19 diagnosis at early stage based on smartwatches and machine learning techniques," *IEEE Access*, vol. 9, pp. 119 476–119 491, Aug. 2021.
- [245] H. S. Maghded, K. Z. Ghafoor, A. S. Sadiq, K. Curran, D. B. Rawat, and K. Rabie, "A novel ai-enabled framework to diagnose coronavirus covid-19 using smartphone embedded sensors: Design study," in *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, Aug. 2020, pp. 180–187.
- [246] A. N. Belkacem, S. Ouhbi, A. Lakas, E. Benkhelifa, and C. Chen, "End-to-end ai-based point-of-care diagnosis system for classifying respiratory illnesses and early detection of covid-19: A theoretical framework," *Frontiers in Medicine*, vol. 8, p. 372, 2021.
- [247] X. Li, C. Li, and D. Zhu, "Covid-mobilexpert: On-device covid-19 patient triage and follow-up using chest x-rays," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2020, pp. 1063–1067.
- [248] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, and M. Nabeel, "Ai4covid-19: Ai enabled preliminary diagnosis for covid-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, p. 100378, Jan. 2020.
- [249] J. Schobel, R. Pryss, M. Schickler, and M. Reichert, "Towards flexible mobile data collection in healthcare," in *2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)*, Jun. 2016, pp. 181–182.
- [250] J. Schobel, R. Pryss, W. Schlee, T. Probst, D. Gebhardt, M. Schickler, and M. Reichert, "Development of mobile data collection applications by domain experts: Experimental results from a usability study," in *Advanced Information Systems Engineering*, ser. Lecture Notes in Computer Science, E. Dubois and K. Pohl, Eds., 2017, pp. 60–75.
- [251] R. Pryss, J. Schobel, and M. Reichert, "Requirements for a flexible and generic api enabling mobile crowdsensing mhealth applications," in *2018 4th International Workshop on Requirements Engineering for Self-Adaptive, Collaborative, and Cyber Physical Systems (RESACS)*, aug 2018, pp. 24–31.
- [252] C. Vogel, R. Pryss, J. Schobel, W. Schlee, and F. Beierle, "Developing apps for researching the covid-19 pandemic with the trackyourhealth platform," in *2021 IEEE/ACM 8th International Conference on Mobile Software Engineering and Systems (MobileSoft)*, 2021, pp. 65–68.
- [253] F. Beierle, J. Schobel, C. Vogel, J. Allgaier, L. Mulansky, F. Haug, J. Haug, W. Schlee, M. Holfelder, M. Stach, M. Schickler, H. Baumeister, C. Cohrdes, J. Deckert, L. Deserno, J.-S. Edler, F. A. Eichner, H. Greger, G. Hein, P. Heuschmann, D. John, H. A. Kestler, D. Krefting, B. Langguth, P. Meybohm, T. Probst, M. Reichert, M. Romanos, S. Störk, Y. Terhorst, M. Weiß, and R. Pryss, "Corona healthstudy- and sensor-based mobile app platform exploring aspects of the covid-19 pandemic," *International Journal of Environmental Research and Public Health*, vol. 18, no. 14, p. 7395, Jan. 2021.

- [254] C. Vogel, J. Schobel, W. Schlee, M. Engelke, and R. Pryss, "Uniti mobile-apps for a large-scale european study on tinnitus," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, Nov. 2021, pp. 2358–2362.
- [255] M. G. Marmot, "Status syndrome challenge to medicine," *JAMA*, vol. 295, no. 11, pp. 1304–1307, Mar. 2006.
- [256] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: Current state and future challenges," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 32–39, Nov. 2011.
- [257] R. Pryss, "Mobile crowdsensing in healthcare scenarios: Taxonomy, conceptual pillars, smart mobile crowdsensing services," in *Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics*, ser. Studies in Neuroscience, Psychology and Behavioral Economics, H. Baumeister and C. Montag, Eds. Springer International Publishing, 2019, pp. 221–234.
- [258] R. Kraft, W. Schlee, M. Stach, M. Reichert, B. Langguth, H. Baumeister, T. Probst, R. Hannemann, and R. Pryss, "Combining mobile crowdsensing and ecological momentary assessments in the healthcare domain," *Frontiers in Neuroscience*, vol. 14, no. 164, 2020.
- [259] A. A. Stone and S. Shiffman, "Ecological momentary assessment (ema) in behavioral medicine," *Annals of Behavioral Medicine*, vol. 16, no. 3, pp. 199–202, Jan. 1994.
- [260] M. Holfelder, L. Mulansky, W. Schlee, H. Baumeister, J. Schobel, H. Greger, A. Hoff, and R. Pryss, "Medical device regulation efforts for mhealth apps during the covid-19 pandemic: An experience report of corona check and corona health," *J*, vol. 4, no. 2, pp. 206–222, Jun. 2021.
- [261] J. Hasell, E. Mathieu, D. Beltekian, B. Macdonald, C. Giattino, E. Ortiz-Ospina, M. Roser, and H. Ritchie, "A cross-country database of covid-19 testing," *Scientific Data*, vol. 7, no. 1, p. 345, Oct. 2020.
- [262] F. Beierle, V. T. Tran, M. Allemand, P. Neff, W. Schlee, T. Probst, J. Zimmermann, and R. Pryss, "What data are smartphone users willing to share with researchers?" *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 2277–2289, 2020.
- [263] N. C. Benda, L. T. Das, E. L. Abramson, K. Blackburn, A. Thoman, R. Kaushal, Y. Zhang, and J. S. Ancker, "'how did you get to this number?' stakeholder needs for implementing predictive analytics: a pre-implementation qualitative study," *Journal of the American Medical Informatics Association*, vol. 27, no. 5, pp. 709–716, 2020.
- [264] S. Benjamens, P. Dhunoo, and B. Meskó, "The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–8, 2020.
- [265] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, no. 8, pp. 500–510, 2018.
- [266] C. Lee, P. Nagy, and S. Weaver, "Cognitive and system factors contributing to diagnostic errors in radiology," *American Journal of Roentgenology*, vol. 201, no. 3, pp. 611–7, 2013.
- [267] E. P. Iyawe, B. M. Idowu, and O. J. Omoleye, "Radiology subspecialisation in africa: A review of the current status," *SA Journal of Radiology*, vol. 25, no. 1, pp. 1–7, 2021.

- [268] D. Dov, S. Z. Kovalsky, J. Cohen, D. E. Range, R. Henao, and L. Carin, "Thyroid cancer malignancy prediction from whole slide cytopathology images," in *Machine Learning for Healthcare Conference*. PMLR, 2019, pp. 553–570.
- [269] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [270] K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and P. A. Friedman, "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management," *Nature Reviews Cardiology*, vol. 18, no. 7, pp. 465–478, 2021.
- [271] R. L. Draeos, J. E. Ezekian, F. Zhuang, M. E. Moya-Mendez, Z. Zhang, M. B. Rosamilia, P. K. Manivanan, R. Henao, and A. P. Landstrom, "Genesis: Gene-specific machine learning models for variants of uncertain significance found in catecholaminergic polymorphic ventricular tachycardia and long qt syndrome-associated genes," *Circulation: Arrhythmia and Electrophysiology*, vol. 15, no. 4, p. e010326, 2022.
- [272] J. A. González-Nóvoa, L. Busto, J. J. Rodríguez-Andina, J. Fariña, M. Segura, V. Gómez, D. Vila, and C. Veiga, "Using explainable machine learning to improve intensive care unit alarm systems," *Sensors*, vol. 21, no. 21, p. 7125, 2021.
- [273] A. Echle, N. T. Rindtorff, T. J. Brinker, T. Luedde, A. T. Pearson, and J. N. Kather, "Deep learning in cancer pathology: a new generation of clinical biomarkers," *British journal of cancer*, vol. 124, no. 4, pp. 686–696, 2021.
- [274] M. Taghiakbari, Y. Mori, and D. von Renteln, "Artificial intelligence-assisted colonoscopy: A review of current state of practice and research," *World Journal of Gastroenterology*, vol. 27, no. 47, p. 8103, 2021.
- [275] K. Ćosić, S. Popović, M. Šarlija, I. Kesedžić, M. Gambiraža, B. Dropuljić, I. Mijić, N. Henigsberg, and T. Jovanovic, "Ai-based prediction and prevention of psychological and behavioral changes in ex-covid-19 patients," *Frontiers in psychology*, vol. 12, 2021.
- [276] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study," *PLoS medicine*, vol. 15, no. 11, p. e1002683, 2018.
- [277] L. G. McCoy, C. T. Brenna, S. S. Chen, K. Vold, and S. Das, "Believing in black boxes: Machine learning for healthcare does not need explainability to be evidence-based," *Journal of clinical epidemiology*, vol. 142, pp. 252–257, 2022.
- [278] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- [279] S. Chakrobarty and O. El-Gayar, "Explainable artificial intelligence in the medical domain: A systematic review," 2021.
- [280] J. D. Fuhrman, N. Gorre, Q. Hu, H. Li, I. El Naqa, and M. L. Giger, "A review of explainable and interpretable ai with applications in covid-19 imaging," *Medical Physics*, vol. 49, no. 1, pp. 1–14, 2021.

- [281] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*, vol. 6, no. 6, 2020.
- [282] K. Hauser, A. Kurz, S. Hagggenmüller, R. C. Maron, C. von Kalle, J. S. Utikal, F. Meier, S. Hobelsberger, F. F. Gellrich, M. Sergon, A. Hauschild, L. E. French, L. Heinzerling, J. G. Schlager, K. Ghoreschi, M. Schlaak, F. J. Hilke, G. Poch, H. Kutzner, C. Berking, M. V. Heppt, M. Erdmann, S. Haferkamp, D. Schadendorf, W. Sondermann, M. Goebeler, B. Schilling, J. N. Kather, S. Fröhling, D. B. Lipka, A. Hekler, E. Krieghoff-Henning, and T. J. Brinker, "Explainable artificial intelligence in skin cancer recognition: A systematic review," *European Journal of Cancer*, vol. 167, pp. 54–69, 2022.
- [283] Y. Zhang, Y. Weng, and J. Lund, "Applications of explainable artificial intelligence in diagnosis and surgery," *Diagnostics*, vol. 12, no. 2, 2022.
- [284] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis," *Medical image analysis*, vol. 79, p. 102470, 2022.
- [285] A. M. Antoniadis, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, and C. Mooney, "Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review," *Applied Sciences*, vol. 11, no. 11, p. 5088, 2021.
- [286] T. P. Quinn, S. Jacobs, M. Senadeera, V. Le, and S. Coghlan, "The three ghosts of medical ai: Can the black-box present deliver?" *Artificial intelligence in medicine*, vol. 124, p. 102158, 2022.
- [287] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.
- [288] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, and A. Holzinger, "Explainable artificial intelligence: Concepts, applications, research challenges and visions," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 2020, pp. 1–16.
- [289] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 150–158.
- [290] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1721–1730.
- [291] R. L. Draelos and L. Carin, "Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks," *arXiv e-prints*, pp. arXiv–2011, 2020.
- [292] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv e-prints*, pp. arXiv–1702, 2017.
- [293] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

- [294] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.
- [295] S. O. Arık and T. Pfister, "Tabnet: Attentive interpretable tabular learning," *arXiv*, 2020.
- [296] S. You, D. Ding, K. Canini, J. Pfeifer, and M. R. Gupta, "Deep lattice networks and partial monotonic functions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 2985–2993.
- [297] R. Confalonieri, T. Weyde, T. R. Besold, and F. M. del Prado Martín, "Using ontologies to enhance human understandability of global post-hoc explanations of black-box models," *Artificial Intelligence*, vol. 296, p. 103471, 2021.
- [298] D. Balduzzi, M. Frean, L. Leary, J. Lewis, K. W.-D. Ma, and B. McWilliams, "The shattered gradients problem: If resnets are the answer, then what is the question?" *arXiv preprint arXiv:1702.08591*, 2017.
- [299] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *ICLR*, 2014.
- [300] P. Lipton, "Contrastive explanation," *Royal Institute of Philosophy Supplement*, vol. 27, p. 247–266, 1990.
- [301] J. A. Recio-García, H. Parejas-Llanovarced, M. G. Orozco-del Castillo, and E. E. Brito-Borges, "A case-based approach for the selection of explanation algorithms in image classification," in *International Conference on Case-Based Reasoning*. Springer, 2021, pp. 186–200.
- [302] R. L. Draelos and L. Carin, "Explainable multiple abnormality classification of chest ct volumes," *Artificial Intelligence in Medicine*, vol. 132, p. 102372, 2022.
- [303] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan *et al.*, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *Systematic reviews*, vol. 10, no. 1, pp. 1–11, 2021.
- [304] M. E. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. Reajul Islam, M. Salman Khan, A. Iqbal, N. Al-Emadi *et al.*, "Can ai help in screening viral and covid-19 pneumonia?" *arXiv e-prints*, pp. arXiv–2003, 2020.
- [305] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.
- [306] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
- [307] A. Adhikari, E. Wenink, J. van der Waa, C. Bouter, I. Tolios, and S. Raaijmakers, "Towards fair explainable ai: a standardized ontology for mapping xai solutions to use cases, explanations, and ai systems," in *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, 2022, pp. 562–568.

- [308] A. Azarpanah, J. Bielby, K. Ingram, E. Mousavi, H. Nye, G. Rockwell, J. Wang, and T. Yoldas, "On the ethics of artificial intelligence," in *CSDH-SCHN 2020*, 2020.
- [309] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021.
- [310] S. Kunjan, T. S. Grummett, K. J. Pope, D. M. Powers, S. P. Fitzgibbon, T. Bastiampillai, M. Battersby, and T. W. Lewis, "The necessity of leave one subject out (loso) cross validation for eeg disease diagnosis," in *Brain Informatics: 14th International Conference, BI 2021, Virtual Event, September 17–19, 2021, Proceedings 14*. Springer, 2021, pp. 558–567.
- [311] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.
- [312] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth *et al.*, "Crisp-dm 1.0: Step-by-step data mining guide," *SPSS inc*, vol. 9, no. 13, pp. 1–73, 2000.
- [313] C. Vogel, J. Schobel, W. Schlee, M. Engelke, and R. Pryss, "Uniti mobile—emi-apps for a large-scale european study on tinnitus," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 2358–2362.
- [314] M. Schleicher, V. Unnikrishnan, P. Neff, J. Simoes, T. Probst, R. Pryss, W. Schlee, and M. Spiliopoulou, "Understanding adherence to the recording of ecological momentary assessments in the example of tinnitus monitoring," *Scientific Reports*, vol. 10, no. 1, p. 22459, 2020.
- [315] P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, no. 1, pp. 1–11, 1968.
- [316] S. Geisser, "The predictive sample reuse method with applications," *Journal of the American Statistical Association*, vol. 70, no. 350, pp. 320–328, 1975.
- [317] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *The Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.
- [318] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation." *Encyclopedia of database systems*, vol. 5, pp. 532–538, 2009.
- [319] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data," *Ecological Modelling*, vol. 406, pp. 109–120, 2019.
- [320] J. Shao, "Linear model selection by cross-validation," *Journal of the American statistical Association*, vol. 88, no. 422, pp. 486–494, 1993.
- [321] H. Meyer, C. Reudenbach, T. Hengl, M. Katurji, and T. Nauss, "Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation," *Environmental Modelling & Software*, vol. 101, pp. 1–9, 2018.
- [322] M. Holfelder, L. Mulansky, W. Schlee, H. Baumeister, J. Schobel, H. Greger, A. Hoff, and R. Pryss, "Medical device regulation efforts for mhealth apps during the covid-19 pandemic—an experience report of corona check and corona health," *J*, vol. 4, no. 2, pp. 206–222, 2021.

- [323] W. Schlee, S. Schoisswohl, S. Staudinger, A. Schiller, A. Lehner, B. Langguth, M. Schecklmann, J. Simoes, P. Neff, S. C. Marcum *et al.*, "Towards a unification of treatments and interventions for tinnitus patients: The eu research and innovation action uniti," *Progress in brain research*, vol. 260, pp. 441–451, 2021.
- [324] F. Beierle, J. Allgaier, C. Stupp, T. Keil, W. Schlee, J. Schobel, C. Vogel, F. Haug, J. Haug, M. Holfelder *et al.*, "Self-assessment of having covid-19 with the corona check mhealth app," *IEEE Journal of Biomedical and Health Informatics*, 2023.
- [325] B. Wetzel *et al.*, "'How come you don't call me?' Smartphone communication app usage as an indicator of loneliness and social well-being across the adult lifespan during the COVID-19 pandemic," *International Journal of Environmental Research and Public Health*, vol. 18, no. 12, p. 6212, 2021.
- [326] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The phq-9: validity of a brief depression severity measure," *Journal of general internal medicine*, vol. 16, no. 9, pp. 606–613, 2001.
- [327] S. Cohen, T. Kamarck, R. Mermelstein *et al.*, "Perceived stress scale," *Measuring stress: A guide for health and social scientists*, vol. 10, no. 2, pp. 1–2, 1994.
- [328] W. Schlee, D. A. Hall, B. Canlon, R. F. Cima, E. de Kleine, F. Hauck, A. Huber, S. Gallus, T. Kleinjung, T. Kypraios *et al.*, "Innovations in doctoral training and research on tinnitus: The european school on interdisciplinary tinnitus research (esit) perspective," *Frontiers in aging neuroscience*, vol. 9, p. 447, 2018.
- [329] E. Rahm, H. H. Do *et al.*, "Data cleaning: Problems and current approaches," *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.
- [330] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.
- [331] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.
- [332] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 149–159.
- [333] P. Lee, S. Bubeck, and J. Petro, "Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine," *New England Journal of Medicine*, vol. 388, no. 13, pp. 1233–1239, 2023.
- [334] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.
- [335] E. Commission, "Regulatory framework proposal on artificial intelligence." *Digital Strategy European Commission*, 2022.
- [336] H. B. Harvey and V. Gowda, "How the fda regulates ai," *Academic radiology*, vol. 27, no. 1, pp. 58–61, 2020.

- [337] P. Balthazar, P. Harri, A. Prater, and N. M. Safdar, "Protecting your patients' interests in the era of big data, artificial intelligence, and predictive analytics," *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 580–586, 2018.