Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems

ANJA THIEME and DANIELLE BELGRAVE, Microsoft Research
GAVIN DOHERTY, Trinity College Dublin

High prevalence of mental illness and the need for effective mental health care, combined with recent advances in AI, has led to an increase in explorations of how the field of machine learning (ML) can assist in the detection, diagnosis and treatment of mental health problems. ML techniques can potentially offer new routes for learning patterns of human behavior; identifying mental health symptoms and risk factors; developing predictions about disease progression; and personalizing and optimizing therapies. Despite the potential opportunities for using ML within mental health, this is an emerging research area, and the development of effective ML-enabled applications that are implementable in practice is bound up with an array of complex, interwoven challenges. Aiming to guide future research and identify new directions for advancing development in this important domain, this article presents an introduction to, and a systematic review of, current ML work regarding psycho-socially based mental health conditions from the computing and HCI literature. A quantitative synthesis and qualitative narrative review of 54 papers that were included in the analysis surfaced common trends, gaps, and challenges in this space. Discussing our findings, we (i) reflect on the current state-of-the-art of ML work for mental health, (ii) provide concrete suggestions for a stronger integration of human-centered and multi-disciplinary approaches in research and development, and (iii) invite more consideration of the potentially far-reaching personal, social, and ethical implications that ML models and interventions can have, if they are to find widespread, successful adoption in real-world mental health contexts.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**;

Additional Key Words and Phrases: Mental health, mental illness, machine learning, systematic review, AI applications, ethics, society + AI, interpretability, interaction design, health care, real-world interventions

ACM Reference format:
Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems. *ACM Trans. Comput.-Hum. Interact.* 27, 5, Article 34 (August 2020), 53 pages.
https://doi.org/10.1145/3398069

The research of Gavin Doherty is funded in part by Science Foundation Ireland Grant no. 13/RC/2106 to the Adapt Centre.
Authors' addresses: A. Thieme (corresponding author) and D. Belgrave, Healthcare Intelligence, Microsoft Research, 21 Station Road, CB1 2FB Cambridge, Cambridgeshire, UK; emails: {anthie, dabelgra}@microsoft.com; G. Doherty, School of Computer Science and Statistics, Trinity College Dublin, College Green, Dublin 2, IRL; email: gavin.doherty@tcd.ie.

1073-0516/2020/08-ART34
https://doi.org/10.1145/3398069

1 INTRODUCTION

Increases in the occurrence and global burden of mental illness have made the prevention and treatment of mental disorders a public health priority [90, 91, 204, 207]. A 2017 US report showed that an estimated 46.6 million adults have been affected by a mental illness. This equates to nearly 20% of the US population alone [169]. Responding to the need for more effective mental health services, the role of digital technology for improving access, engagement, and outcomes of therapeutic treatment is increasing in importance and has led to a wide range of health technologies and applications (e.g., [41, 156, 187]). These include mobile apps and wearable devices to assist symptom monitoring and health risk assessments [12, 48], computerized treatments [49, 157, 171], and mental health peer or community support [99, 137, 150]. These systems as well as people's everyday technology interactions and the information that is accumulated in electronic health care records (EHR) increasingly provide a wealth of personal health and behavioral data [65, 116, 124]. Eyre et al. [56] even suggest that the field of *"mental health captures arguably the largest amount of data of any medical specialty"* (p. 21). Growth in data availability alongside improvements to computing power has led to a surge in research and applications of machine learning (ML) technologies [25, 186]. The field of ML extends statistical and computational methods to construct more robust systems with an ability to automatically learn from data [173]. These techniques have been applied successfully in the domains of gaming and recommender systems and show promise in helping to understand large-scale health data. By offering new routes to improving our understanding of human behaviors and predicting or optimizing outcomes [85, 173], ML approaches are increasingly being explored for mental health (e.g., [40, 42, 83, 166]).

In recent years, reviews of the literature and research surveys that focus on applications of ML for mental health have started to emerge in the medical and clinical psychology domain. Existing research assesses the accuracy, reliability, and effectiveness of algorithms [100, 158], as well as opportunities and challenges for their adoption in practice [25, 124]. Much of the work addresses algorithm use in the area of neuroscience, specifically in neuroimaging research (e.g., [10, 173, 181, 206]). Other works study algorithmic performance in predicting the outcomes of clinical interventions (e.g., pharmacological treatments) for specific mental health conditions (e.g., [100]), or discuss approaches to identify key behavioral markers for clinical states from mobile mental health sensing data [124, 158]. To provide a better overview of the different application areas of ML in the mental health domain, Shatte et al. [173] recently conducted a scoping review to map the key concepts underpinning this field from 300 literature records. The authors identified four main application domains, with the majority of studies investigating the following: (i) detection and diagnosis of mental health conditions, and others addressing (ii) prognosis, treatment, and support, (iii) public health, or (iv) research and administration. They conclude that, by generating new insights into mental health and well-being, these works demonstrate the potential of ML to improve the efficiency of clinical and research practices.

The impact of ML in mental health will be strongly mediated by the design of systems which employ ML, which motivates us to examine recent research in computing and HCI addressing this topic. Complementing research perspectives from medical science and clinical psychology, our article presents a systematic review of the ACM Guide to Computing Literature to derive a deeper understanding of the current landscape of ML applications for mental health from an HCI and computing science perspective. In this regard, our work builds on a recent review by Sanches et al. [170], which mapped the design space of technologies for supporting affective health as reported in HCI; and identified that most innovation has occurred in the areas of automated diagnosis, and self-tracking. As researchers who are actively working at the intersection of HCI, ML and mental health, we are excited about the prospective benefits that ML techniques could bring to mental health. Simultaneously, from the outset of the review, we were also aware that the *development of effective and implementable ML systems is bound up with an array of complex, interwoven socio-technical challenges*. In this regard, our review is likely shaped by both our *cautious optimism that ML approaches can be usefully and successfully applied* in this domain; and a strong *human-centered* perspective on technology development as well as commitment to creating *responsible AI* applications that seek to *improve societal outcomes*. As a result, we take, at times, a slightly more critical view on research that proposes potentially impactful real-world interventions yet remains solely centered on technical innovation. Aiming to move the field forward in achieving many of its ambitious goals for real-world impact, we invite the community to engage more actively and critically with many of the complex challenges involved in order to realize successful use of ML in mental health.

These challenges include *generating large-scale, high-quality datasets*, which are representative of the diversity of the population, and gaining access to such datasets with the purpose of developing more robust and fairer ML models (cf. [20]). Mental health, in particular, affects a broad spectrum of people—spanning different demographics (age, gender, ethnicity), geographic locations, and socio-economic statuses—which calls for the inclusion of a wide range of people for this diversity to be reflected in the dataset to mitigate risks of bias [66, 73, 76, 80]. However, data collection is costly and particularly complicated where information is deeply personal as well as sensitive due to the stigma that is often associated with mental conditions [29, 53, 116, 190]. Subsequently, this raises the question of whether or not people should trust ML applications with the collection and processing of their personal data, and to what extent and by what mechanisms people should agree to the collection of such personal data.

These challenges are further exacerbated by forms of *error, uncertainty, and bias,* which are an obstacle for the ready deployment of "state-of-the-art" ML algorithms into real-world intelligent systems (cf. [190]). Even in cases where good accuracy can be achieved, there is always the challenge of generalization, whereby models that are trained with high accuracy in one scenario may not transfer to scenarios outside of the environment of the training dataset [106]. This may introduce various sources of bias in the model, for example, demographic disparity due to under-representation of certain groups in the training data [19, 66, 73]. Such disparities may become magnified in sensitive domains such as mental health. This brings into question the ethical implications of deploying a ML algorithm into an actionable health diagnosis or treatment recommendation. This needs an interdisciplinary approach to model *interpretability*, where clinical, HCI, and other domain experts support the understanding of uncertainty, accuracy, and potential biases in ML outputs.

Finally, if ML applications are to find widespread adoption and success in real-world mental health contexts, it is also crucial to consider potentially far reaching *personal, societal, and economic implications* that the introduction of ML interventions can have. This includes *ethical questions* about responsibility and accountability for ML-directed decision making [15]; risks of potentially fallible ML outputs and biases; malicious uses of ML (see related works in domains of criminal justice, loan decisions [161] or automated facial analysis [24], and adversarial attacks in image processing [74] and speech recognition [37]); or digital exclusion due to lack of knowledge, access or other barriers to technology use [70].

To provide a knowledge base to inform future research, our analysis of the computing literature presents a quantitative synthesis and qualitative narrative overview [69] of ML applications in mental health. Our aims are to (i) provide a comprehensive introduction to this important and evolving area of research, (ii) highlight existing trends and gaps to guide future work and encourage a stronger involvement by the HCI community, and (iii) sensitize the community to many of the complex technical, societal and ethical challenges that are bound up with the development of ML applications, if they were to be effective and implementable in health care practice. In this regard, our literature review was guided by six main questions:

- What types of ML models and applications are currently being developed for mental health?
- What motivates the use of ML in the reported works and what aspects of mental health do they target?
- What types and scale of data is used for ML analysis and how is access to mental health data achieved?
- What techniques were applied (and challenges encountered) in developing and evaluating ML models?
- What key learnings are reported in the literature and to what extent do they apply to real-world contexts?
- To what extent do the papers describe ethical challenges or implications?

First, we describe our systematic review methodology; followed by the findings that were extracted from the review corpus papers. The article concludes with a critical discussion of the findings and provides a set of concrete suggestions for steps forward in developing ML applications and systems that are useful, ethical and implementable in supporting mental health.

2 METHODOLOGY

To structure the identification and selection of relevant articles for our review, we followed the PRIMSA literature review guidelines [104, 123].

2.1 Record Identification

Relevant papers were identified by searching the electronic database of the Association for Computing Machinery (ACM) Guide to the Computing Literature, which is the most comprehensive bibliographic database in the field of computing and HCI research. It integrates full text-articles of conference proceedings, journals, magazines, books, and abstracts of key publishers including ACM, IEEE, Springer, Elsevier, John Wiley & Sons, and Kluwer. The final corpus presented here resulted from a search conducted on the 15th of November 2019. It included the search terms "mental health" AND "ML" (see full query syntax[1]), which identified 122 records.

2.2 Record Selection

The titles and abstracts of the 122 records were independently screened by two researchers to determine their fit with regards to addressing an application of "ML" in the context of "mental health." Papers were eligible for inclusion if they reported an application of ML for understanding, detecting, diagnosing, treating affective mental health problems or conditions (e.g., stress, depression, anxiety), psycho-social functioning (e.g., general mental health or well-being [188]), and practices to support mental health more broadly (e.g., mental health care providers).

Papers were excluded if they described topics of neuroscience, neurobiology, or neurological conditions—including cell structure, cortex, and (f)MRI research (n = 22 [3, 54, 59, 72, 92, 94, 95, 96, 98, 103, 108, 115, 117, 131, 146, 180, 183, 185, 195, 199, 214, 220]), and in one case epilepsy [5]. We also excluded neurodevelopmental disorders such as autism or ADHD [60, 175] that present primarily as behavioral conditions. Although they both can affect a person's ability to socialize and communicate with others, we focused our review on psychosocially oriented mental health conditions that, instead, are primarily caused or influenced by life experiences, as well as

[1]Full Query Syntax used: "query": {(+"machine learning" +"mental health")} "filter": {"publicationYear":{"gte":1990}}, {owners.owner=GUIDE}.

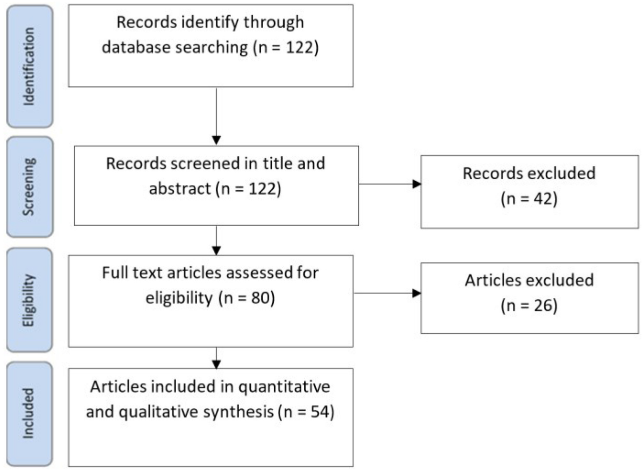


Fig. 1. Procedural flowchart following the PRIMSA guidelines.

maladjusted cognitive and behavioral processes. Among the most common psychosocial conditions are mood, anxiety, eating, personality, and substance abuse conditions as well as schizophrenia. This selection criterion is consistent with the mental health literature and other systematic reviews on affective mental health [29, 170]. Further excluded were papers that described ML research outside a specific focus on mental health (n = 16 [21, 26, 32, 43, 75, 81, 88, 114, 118, 120, 160, 167, 200, 208, 215, 221]); or that otherwise did not fit thematically (n = 20 [11, 14, 18, 29, 30, 36, 55, 87, 101, 107, 110, 113, 138, 143, 147, 174, 178, 194, 197, 198]). Examples include: a workshop on digital biomarkers [55], a study of the effectiveness of eye-movements [198], or encryption methods for protecting the privacy of databases [18]; as well as review or overview papers that did not directly report an application of ML for mental health [29, 30, 194]. Seven records could not be accessed (e.g., [142]). Based on a review of title and abstract, we identified 38 records as eligible, excluded 42 records, and noted uncertainty or disagreements in the classification of 42 records, which required full-text screening. Following full-text review, another 26 papers were excluded, leaving a final corpus of 54 articles for inclusion in the systematic review (Figure 1).

2.3 Data Extraction

To assist the systematic extraction of data from the papers, we created a data extraction sheet (see Table 1 for an excerpt). It includes columns for characterizing the papers by authors, affiliations, title, abstract, publication type and year, and individual columns to describe the following: the type of ML application, its motivation, main data source, and target users. We further recorded information about data access, data subjects, data scale, and data processing steps, the ML algorithms used, and approaches to their evaluation. For each paper, we also summarized the main research insights that were reported and listed any descriptions of ML-specific data challenges (e.g., data quality, bias, fairness, uncertainty/error, algorithmic interpretability). Finally, we noted if the works include topics such as real-world application, study or design challenges, and discuss ethical issues. The extraction sheet was pilot tested on ten randomly selected papers that fulfilled the inclusion criteria. It was first developed and completed by one of the review authors, and then checked by another author. Each paper was analyzed using this template. Once data extraction was completed, we added additional columns to aid synthesis across papers. This included among others: the papers' main contribution type (Figure 2, right), target mental health behavior or condition

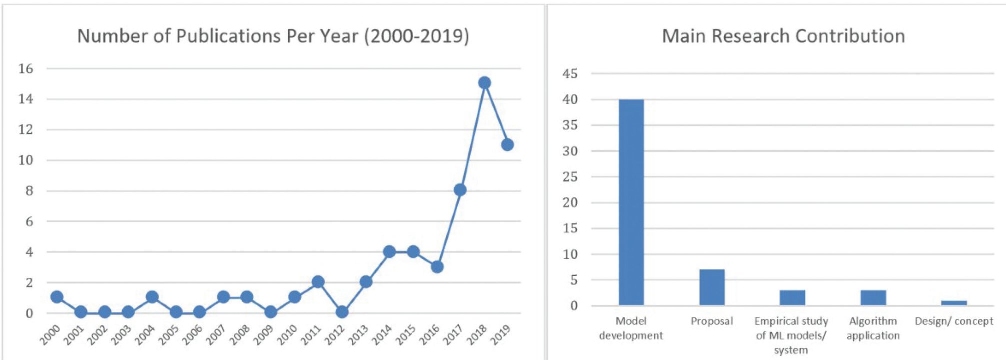


Fig. 2. Left: Graph showing an increasing trend in the number of ML mental health publications over time. Right: Frequency distribution of the different research contribution types of the papers in the review corpus.

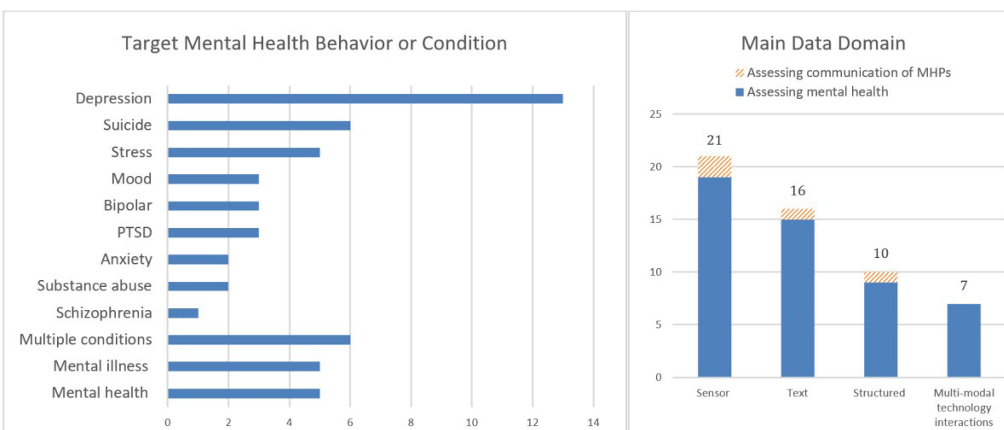


Fig. 3. Left: Distribution of the types of mental health behaviors or conditions that were the target of support across all review papers. Right: Frequency distribution of the main data domains that were used in the respective papers to extract insights for mental health.

(Figure 3, left), and the category of ML algorithm used. The findings provide a quantitative and narrative summary of the corpus with detailed examples of relevant publications. This approach has been chosen to reflect both the breadth and depth of the trends and challenges reported; as well as to help identify any gaps in the literature.

3 FINDINGS

The final review corpus includes papers published between the years of 2000 and 2019. Publications increased in recent years (Figure 2, left), with nearly two-third of all papers published in the last three years.

Of the 54 papers, 33 are conference publications (7 abstracts, 5 short- and 21 full-length proceeding papers), 14 are journal articles, and 7 symposium or workshop papers. Furthermore, Figure 2 (right) shows how the vast majority of the papers primarily describe the development of a ML model based on specific data as their main research contribution (n = 40). Seven papers are proposals of specific concepts [28, 82, 154], data methods [31], models [184], or systems [193, 217], and three apply existing ML algorithms to better understand [209] and assess mental health [201], or improve the communication of mental health providers [205]. Furthermore, few papers describe the conduct of empirical studies of an end-to-end ML system [78, 140] or assess the quality of ML predictions [53]. One paper specifically discusses design implications for user-centric, deployable ML systems [77].

3.1 Types of ML Applications, Their Data and Mental Health Focus

This section describes what mental health behaviors or conditions were targeted, what types of data was used to extract mental health-related insights, and the types of ML applications and models that were developed.

3.1.1 Target Mental Health Behaviors or Conditions. The works reviewed can broadly be grouped into two main application areas: (i) the majority of papers that come under *assisting understanding, detection and diagnosis of mental health status*, (n = 49); and (ii) a small portion of papers *assess patient–clinician relationships* (n = 1) or *seek to improve treatment* (n = 4).

Of all 54 papers, a large proportion described a focus on supporting people with mental health behaviors or conditions of depression (n = 13) and suicide (n = 6). Some works addressed stress (n = 5), bipolar disorder (n = 3), mood (n = 3), PTSD (n = 3), anxiety (n = 2), substance abuse (n = 2), or schizophrenia (n = 1). A number of papers (n = 6) targeted multiple mental health conditions (i.e., schizophrenia and mania [45]); and others focused more broadly on mental illness (n = 5), or mental health (n = 5). See Figure 3 (left) for details.

3.1.2 Main Data Domains for ML in Mental Health. We identified four main types of data (Figure 3, right) that were used to extract mental health-related insights through ML: (i) *sensors*, (ii) *text*, (iii) *structured data*, and (iv) *multi-modal technology interactions*.

Sensor-based ML approaches were most common (n = 21). Here, the majority of papers reported uses of mobile phone sensors for data collection (n = 9) or analyzed audio signals (n = 6 [23, 31, 77, 78, 122, 168]). The second largest data source was text (n = 16), which was primarily extracted from social media (n = 11); and, in a few instances, from SMS [134] or text messaging [205]; and from clinical [2, 45] or suicide notes [145]. Papers that analyzed structured data (n = 10) included the evaluation of questionnaires (n = 7) and health records (n = 3). Several papers (n = 7) described complex multi-modal systems, or frameworks that built on everyday technology [82, 140, 193, 222], robot [154], or human/virtual agent [155, 184] interactions.

3.1.3 Types of ML Applications for Mental Health. Next, we describe, the types of ML applications or models that have been developed in each of the main application areas of (i) understanding, detecting and diagnosis of mental health symptoms or outcomes; and (ii) assessing patient–clinician relationships and improving mental health treatment.

3.1.3.1 Understanding, Detecting and Diagnosis of Mental Health Status. A large proportion of the papers described uses of ML to assist in the *detection or diagnosis of mental health symptoms or conditions* (n = 27). Many of these works focus on the (early) detection (and monitoring) of *depression or its symptoms* (n = 10) [31, 33, 44, 57, 62, 122, 136 154, 211, 222], most often through the analysis of acoustic features of speech [31, 122] or Twitter tweets [33, 86, 211]. Other examples include the detection of mood states from mobile sensing data [128, 176], or phone typing dynamics [27] as well as stress assessments from location [218], biometrical and accelerometer data [67]. This is complemented by recent trends in analyzing human-robot [154] or agent interactions [155] to help assess peoples' mental health status. Furthermore, text analysis was performed to detect and *automatically extract diagnostic information* from written narratives or psychiatric records [45], while questionnaire data was studied to help *differentiate between mental health states or diagnosis* such as patients who experience bipolar I or bipolar II [61].

Aside from mental health detection and diagnosis, a significant proportion of the papers described approaches to *understanding and predicting mental health risks* (n = 8). Predominantly this included efforts to *predict future suicide risks* from either sensor data [6], health records [2, 192], or text [145, 134]. Examples include: the analysis of written suicide notes [145]; of suicidality periods from the SMS messages of individuals with a history of suicidal behaviors [134]; and of suicide risk at time of a person's referral to mental health services [192], and subsequent periods [2]. Outside of suicide prediction, individual papers sought to help predict: *episodes of mania or depression* in people who experience bipolar conditions [50]; *risks of re-hospitalization* of outpatients with severe mental health difficulties [144]; and *experiences of patient stress* [139].

In the context of social media analysis, a number of papers (n = 5) further aimed to better understand *the linguistic characteristics of mental health-related content shared in online communities*; focusing primarily on Reddit[2] posts, and in one case, data from Live Journal[3] [133]. Here, text-mining approaches were used [89] or proposed [28] to *identify helpful and unhelpful comments* in online mental health communities to assist human moderators to prioritize their responses to comments [28, 89]. Saha and De Choudhury [165] further developed a classifier for *inferring expressions of stress from Reddit posts* by college students before and after incidences of gun violence; while others extracted linguistic features and topics in mental health communities to *learn more about themes discussed online* [133, 141].

Outside of these three main categories, more isolated investigations included: the application of ML to gain more insight into *what factors* (e.g., *psychological symptoms, contextual influences) may impact a person's mental health the most* [63, 209] and *their relation to mental health outcomes* [135, 201]. Further, Tsiakas et al. [193] described a prototype system that engages the user in dialogue with a female avatar that asks a series of questions to screen for symptoms of depression and anxiety. Proposed as an adaptive system, the *screening questions are optimized* and encouragements offered based on the users responses and their emotional state. Table 1 provides summaries of all paper records, their purpose and targets; and illustrates how the use of different data domains (e.g., sensor, text) is distributed across the corpus.

Next, we expand on the small number of papers that did not focus on assessing mental health, and instead explored how ML could help assess patient–clinician relationships and improve mental health treatment.

3.1.3.2 Assessing Patient–Clinician Relationships and Improving Mental Health Treatment. Although much of the research that focused on mental health assessment has the motivation to provide effective tools to aid clinicians and other care providers in their work (e.g., [2, 28, 61, 86, 89, 122, 134, 145]), several papers (n = 5) described investigations of *how ML techniques can be leveraged to assess the patient–clinician relationship* and *improve the content or delivery of mental health treatment.*

For example, Aguilar-Ruiz et al. [4] developed a knowledge model from questionnaire data about psychiatric hospital patients' experiences of their relationship with their doctors to help improve doctor communication. The remaining papers either sought to help identify what may

[2]Reddit (www.reddit.com) is a website that offers a collection of forums, where users can share content or comment on other peoples' posts. The service consists of more than one million communities, called 'subreddits', and has more than 330 million monthly active users.
[3]Live Journal (www.livejournal.com) is a social networking service with approximately 30 million monthly visitors. Users have a profile page, can maintain a personal blog, connect and communicate with others, and form an online community in the form of a collective blog [133].

Table 1. Datasheet Excerpt of all 54 Papers Including Data Domain, Purpose and Description of the ML Application or Approach, Its Motivation, and Target Mental Health Symptom or Condition

Reference	Data Domain	Purpose	ML Application/Approach (What)	Motivation (Why)	Mental Health Target
	Sensors (21)				
Chang et al. [31]	Audio	Detecting symptoms/ condition	Development of an automatic mental-health monitor based on the human voice. Initial step: developing categorization of voice utterances for analysis of mental health symptoms.	To assist in the early diagnosis and longitudinal monitoring of mental illness symptoms in everyday speech conversation.	Depression
Broek et al. [23]	Audio	Detecting symptoms/ condition	Development of a speech-based stress indicator. Comparison of controlled storytelling study (ST) with an ecologically valid reliving (RL) study.	To support efficient treatment of PTSD, which requires objective understanding of patients' emotional distress.	PTSD
Salekin et al. [168]	Audio	Detecting symptoms/ condition	Development of a weakly supervised learning framework for detecting social anxiety and depression from long audio clips that includes a novel feature modeling technique (NN2Vec).	To objectively and unobtrusively detect speakers high in social anxiety or depression symptoms that do not require extensive equipment or clinical training.	Anxiety
Mitra et al. [122]	Audio	Detecting symptoms/ condition	Development of a depression-level recognizer based on a set of acoustic features in spoken audio.	To assist accurate diagnosis of depressive symptoms.	Depression
Frogner et al. [62]	Accelerometer	Detecting symptoms/ condition	Development of multiple ML models to detect presence and level of depression from motor activity recordings.	To accurately detect depression from very easy to obtain motor activity.	Depression
Mallol-Ragolta et al. [112]	Body (skin conductance)	Detecting symptoms/ condition	Development of a multi-modal approach to estimate changes in PTSD symptom severity based on self-reports and skin conductance physiology.	To aid non-intrusive measures of PTSD symptom severity through skin conductance responses; reducing need for self-report.	PTSD
Rabbi et al. [153]	Multiple (audio + activity)	Detecting symptoms/ condition	Development + study of multi-modal mobile sensing system to simultaneously assess mental and physical health from passive sensing of everyday speech in naturalistic conditions.	To continuously monitor a person's mental well-being via mobile sensing that is easy, low cost, secure + protects privacy.	Mental health (generic)
Gjoreski et al. [67]	Multiple (body)	Detecting symptoms/ condition	Development of a method for continuous detection of stressful events from a commercial wrist worn device.	To assist mental health and well-being self-managing by developing a stress-detection application as part of a mobile app.	Stress
DeMasi and Recht [44]	Mobile phone (GPS)	Detecting symptoms/ condition	Modeling the relationship between user characteristics and algorithmic predictions of peoples' daily mental well-being from smartphone GPS data.	To explore if mental well-being can be inferred from smartphone behavioral data and automatically tracked over time.	Depression
Zakaria et al. [217]	Mobile phone/laptop (Wi-Fi)	Detecting symptoms/ condition	Proposed development of a stress monitoring system that is driven by indoor localization technology to predict excessive stress.	To automatically and non-intrusively detect signs of excessive stress from mobile phone without the need for installing an app.	Stress
Zakaria et al. [218]	Mobile phone/laptop (Wi-Fi)	Detecting symptoms/ condition	Development of StressMon, a stress and depression detection system that leverages location data from a university Wi-Fi system to better understand physical social interactions.	To help detect individuals' stress and depression early and overcome need for app use.	Multiple: Stress + depression
Cao et al. [27]	Mobile phone (acceleration)	Detecting symptoms/ condition	Development of an architecture for modeling mobile phone typing dynamics for inferring mood states in bipolar patients (based on a late fusion strategy for data integration).	To assist unobtrusive detection of psychiatric diseases in patient's daily lives.	Bipolar

(Continued)

Table 1. Continued

Reference	Data Domain	Purpose	ML Application/Approach (What)	Motivation (Why)	Mental Health Target
Quisel et al. [152]	Mobile phone (multiple)	Detecting symptoms/ condition	Testing pre-existing classifier of varied self-reported mental health and nervous system conditions (multi-task trained CNN model) for different data collection time windows.	To identify effective (least disruptive) time window for passively collected mobile-health data with high accuracy.	Multiple conditions
Spathis et al. [176]	Mobile phone (multiple)	Detecting symptoms/ condition	Development of ML models to predict mood from passive mobile phone sensing data and personality trait questionnaire responses.	To accurately predict mood from passive data for mental health assessment to avoid frequent experience sampling (burden).	Mood
Morshed et al. [128]	Mobile phone (multiple)	Detecting symptoms/ condition	Development of ML models to predict mood instabilities from passive sensing/multi-modal data in situated communities.	To develop a passive method to model mood states at scale.	Mood
Wang et al. [201]	Mobile phone (multiple)	Understanding mental health	Development of the StudentLife smartphone app that incorporates sensing + EMA to assess college student mental health, academic performance + behavioral trends.	To unobtrusively capture student life via objective smartphone data to understand mental health + education outcomes.	Mental health (generic)
Nosakhare and Picard [135]	Mobile phone (multiple) + activity	Understanding mental health	Development of framework to map multi-modal behavioral observational data to meaningful feature representations, and to uncover behavior patterns predictive of stress/well-being.	To provide tools for objective data analysis to help individuals monitor their well-being using real-world measurements.	Stress
Doryab et al. [50]	Mobile phone (multiple)	Understanding/ predicting risks	Development of a method to infer the progression of a primary health parameter and applying parameter ranking to see which behavioral data has the highest "impact" on health.	To assist prediction, prevention and general self-management of episodes of mania and depression of people with bipolar.	Bipolar
Alam et al. [6]	Multiple (body)	Understanding/ predicting risks	Development of a cloud-based system architecture for collecting and processing real-time body-sensor data as well as additional patient information for assessing suicide risks.	To effectively predict (normal, atypical, and suicidal) mental states of patients with mental health conditions to monitor suicide risk.	Suicide
Hirsch et al. [77]	Audio (counselling session)	Improving treatment	Design considerations in developing ML system to automatically assess motivational Interviewing (MI) skills of psychotherapists from audio recordings of counselling session.	To effectively assess therapist performance to aid their skills development and retention for better patient outcomes.	Substance abuse
Hirsch et al. [78]	Audio (counselling session)	Improving treatment	User study of a ML system to automatically assess the motivational Interviewing (MI) skills of psychotherapists directly from the audio recording of a counselling session.	To effectively assess therapist performance to aid their skills development and retention for better patient outcomes.	Substance abuse
	Text (16)				
Chancellor [28]	Social media: Reddit	Understanding mental health content	Development of statistical methods to identify "helpful" vs. "unhelpful" online mental health/wellness comments.	To understand deviant behaviors on online mental health communities.	Multiple: Eating disorder + suicide
Saha and De Choudhury [165]	Social media: Reddit	Understanding mental health content	Development of a ML classifier for inferring expressions of stress from social media posts and time series analysis to examine temporal patterns (before/after) gun violence.	To study the expression of stress in social media in colleges affected by gun-violence incidents.	Stress
Kavuluru et al. [89]	Social media: Reddit	Understanding mental health content	Development of identifiers of "helpful" comments posted within the Reddit community: Suicide Watch (SW), using varied text-mining techniques.	To assist human moderators who review online posts through indicating and/or prioritizing useful/helpful comments.	Suicide

(Continued)

Table 1. Continued

Reference	Data Domain	Purpose	ML Application/Approach (What)	Motivation (Why)	Mental Health Target
Park et al. [141]	Social media: Reddit	Understanding mental health content	Application of methods of text mining, qualitative analysis and data visualization to compare discussion topics in three different online mental health communities on Reddit.	To inform the future design of mental health related online communities and patient education programs.	Multiple
Nguyen et al. [133]	Social media: Live Journal	Understanding mental health content	Application of text-mining to better understand linguistic features and topics related to mental health discussed within online communities on the Live Journal platform.	To improve understanding of mental illnesses.	Depression
Fatima et al. [57]	Social media: Live Journal	Detecting symptoms/ condition	Development of three ML models for classifying depressive posts, communities and the degree of depression from online social media (Live Journaling posts).	To make use of user-generated content to identify depression and characterize its degree of severity.	Depression
Gaur et al. [64]	Social media: Reddit	Detecting symptoms/ condition	Development of multi-class classification algorithm that analysis mental health subreddit posts and quantifies their relationship to DSM-5 categories.	To cost-effectively offer actionable information to clinicians about a patients' mental health for web-based intervention.	Mental illness (generic)
Joshi et al. [86]	Social media: Twitter	Detecting symptoms/ condition	Development of a model to identify different types of mental health conditions from peoples' social media tweets.	To help early diagnosis of mental illness to facilitate help seeking from professional counselors (in India).	Mental illness (generic)
Yazdavar et al. [211]	Social media: Twitter	Detecting symptoms/ condition	Development of a statistical model for monitoring different symptoms of depression by modeling user-generated content in social media tweets over time.	To unobtrusively monitor clinical depressive symptoms in social media.	Depression
Chen et al. [33]	Social media: Twitter	Detecting symptoms/ condition	Development of a model that includes measures of eight basic emotions and temporal data as features in prediction self-reported diagnosis of depression on Twitter.	To earlier identify and better monitor people with, or at risk of depression, from Twitter.	Depression
Ernala et al. [53]	Social media: Twitter + Facebook	Detecting symptoms/ condition	Empirical study to assess internal and external predictive validity of different social media-derived proxy diagnostic signals for schizophrenia.	To obtain clinically valid diagnostic information from sensitive patient populations.	Schizophrenia
Diedrich et al. [45]	Stories + psychiatric reports record	Detecting symptoms/ condition	Development of an ML model to determine schizophrenia from written text narratives; and use of clustering techniques to extract key diagnostic categories from psychiatric reports.	To determine mental health problems through text classification and achieve more accurate diagnostic classification systems.	Multiple
Nobles et al. [134]	Messages (SMS)	Understanding/ predicting risks	Development of a model that identifies periods of suicidality. Report on collection + analysis of text messages of individuals with a history of suicidal behaviors.	To identify subtle clues in text communication as indicators of heightened suicide risk for more effective prevention.	Suicide
Pestian et al. [145]	Suicide notes	Understanding/ predicting risks	Development of a classifier for predicting suicide through natural language processing of written suicide notes.	To provide emergency departments with an evidence-based risk assessment tool for predicting repeated suicide attempts.	Suicide
Adamou et al. [2]	Medical notes (from Health record)	Understanding/ predicting risks	Application of text-mining techniques of medical notes to improve accuracy of a predictive model of suicide risk within 3 or 6 months at point of referral to mental health services.	To increase accuracy of predictive model in efforts to provide a tool that could support clinical assessment of suicide risk.	Suicide

(Continued)

Table 1. Continued

Reference	Data Domain	Purpose	ML Application/Approach (What)	Motivation (Why)	Mental Health Target
Wilbourne et al. [205]	Messages (chat app)	Improving treatment	Use of ML tools to aid supporters of text-based, technology-enabled mental health intervention to assess the quality of their coaching in real-time.	To evaluate and improve the quality of the responses that Silby coaches provide.	Mental health (generic)
	Structured Data (10)				
Galiatsatos et al. [63]	Questionnaire (from Health record)	Understanding mental health	Development of Bayesian models to better understand the most significant psychological symptoms in mental health patients with depression.	To better understand the kinds of factors that affect mental health patients who have thoughts of death or suicide.	Depression
Feng et al. [61]	Questionnaire (from Health record)	Detecting symptoms/ condition	Development of a classifier to distinguish bipolar I from bipolar II patients using only a small number of features.	To more conveniently, efficiently, and accurately distinguish between bipolar I and II assessment.	Bipolar
Srividya et al. [179]	Questionnaire	Detecting symptoms/ condition	Application of clustering for data labelling and subsequent development of a classifier to determine the mental health state of a person as mentally stressed, neutral or happy.	To identify individuals who are mentally distressed to support early detection, and thereby, to benefit society.	Mental health (generic)
Spathis et al. [177]	Questionnaire	Detecting symptoms/ condition	Development of multi-task encoder-decoder RNN that learns patterns from different users to predict their mood from a limited number of self-reports	To provide an effective, ready-to-use tool for early diagnosis of mood issues at scale via mobile mental health apps.	Mood
Ojeme and Mbogho [136]	Health record	Detecting symptoms/ condition	Development of a class-bridge multi-dimensional Bayes network classification approach to simultaneously identify depression and physical illness.	To provide reliable and clinician interpretable diagnostic results for detection of depression + physical illness in Nigeria.	Depression
Yang and Bath [209]	Questionnaire	Understanding mental health	Application of 5 ML models and their combinations to better predict and understand factors of depression in older people.	To improve understanding of underlying pathophysiology of depression for developing appropriate interventions.	Depression
Panagiotako-poulos et al. [139]	Questionnaire	Understanding/ predicting risks	Development of an application for archiving and retrieving patient health records. Data analysis to find associations in context data and to predict patient stress in a given context.	To provide medical staff applications that make use of multi-parameter contextual data collected over longer-term periods.	Anxiety
Patterson and Cloud [144]	Health record	Understanding/ predicting risks	Application of artificial neural networks (ANNs) for predicting re-hospitalization of severely mentally ill outpatients.	To develop + deploy systematic risk assessment decision support tool to guide intervention; reducing rates + costs of rehospitalization.	Multiple
Tran et al. [192]	Health record	Understanding/ predicting risks	Development of a framework to automatically predict low-, moderate-, and high-risk of suicide given mental health history, risk assessment and clinical intervention data.	To improve early detection of suicide and prevention.	Suicide
Aguilar-Ruiz et al. [4]	Questionnaire	Assessing patient–clinician relationship	Development of knowledge model for describing the relationship between (psychiatric) patients and their doctors in a hospital context.	To provide insight that would enable doctors to better communicate with their patients to increase patient satisfaction.	Mental illness (generic)
	Multi-modal system use (7)				
Jain and Agarwal [82]	Chatbot, web media activity + wearables	Detecting symptoms/ condition	Development of a methodological framework for creating an electronic health portfolio based on daily computer interactions for psychiatric symptom diagnosis + prognosis.	To help early diagnosis of mental illness to facilitate help seeking, share health progression, and optimize treatments.	Mental illness (generic)

(Continued)

Table 1. Continued

Reference	Data Domain	Purpose	ML Application/Approach (What)	Motivation (Why)	Mental Health Target
Tavabi [184]	Embodied agent (video, audio, text)	Detecting symptoms/ condition	Proposed development of multi-modal ML methods for augmenting embodied interactive agents with emotional intelligence and assist in mental health assessment.	To augment clinical resources in diagnosis and treatment of patients through automatic behavior analysis.	Mental illness (generic)
Zhou et al. [222]	Interaction + video data, questionnaires	Detecting symptoms/ condition	Development of a multimodal signal system that analysis a person's social media stream and images of a close-up video (i.e. from mobile) to monitor and predict mental health states.	To develop effective, physically noninvasive, low-cost approach to assess mental health via pervasive multimodal sensors.	Depression
Rastogi et al. [154]	Multi-modal robot Interactions	Detecting symptoms/ condition	Development of a CBT-based, multi-modal, humanoid robot interaction framework for depression detection.	To study signs of depression from "unobtrusive" multi-modal communication with social robot.	Depression
Ray et al. [155]	Multi-modal human/agent interaction data	Detecting symptoms/ condition	Development of a novel ML framework with attention mechanisms at several layers to identify + extract important features from multi-modal data to predict level of depression.	To use behavioral cues to predict depression severity to address subjectivity problems of existing diagnostic tests.	Depression
Tsiakas et al. [193]	Audio-visual + structured data	Optimizing health screening	Development of dialogue system that models optimal transitions in a screening process for anxiety and depression based on user response to questions + emotions.	To create adaptive dialogue to aid effective symptom screening and provide referrals to relevant treatment resources.	PTSD
Paredes et al. [140]	Phone app use data; user traits + self-reports	Improving treatment	User study of a smart-phone application that uses ML for personalized recommendations of constructive stress coping behaviors and services.	To help people cope better with stress at scale, beyond what individual or group therapies can provide today.	Stress

be the optimal treatment intervention for a particular individual [140] or help improve the communication skills of mental health care professionals (MHPs) as part of talk-based psychotherapy interventions [77, 78, 205]. For instance, Paredes et al. [140] applied ML in a mobile phone app to help *recommend personalized coping strategies* for stress management. Their system learned from users' engagement with different stress interventions to predict *which intervention—out of a given set—may be correlated with stress reduction for a particular person*, which becomes the basis for personalized intervention recommendations.

In contrast, the other three papers focus on ways to *improve the treatment itself by assisting MHPs to improve their professional communications.* Hirsch et al. [77, 78] describe the design of an assessment and training tool for counsellors that uses speech and language processing to *automatically generate evaluations of the motivational interviewing (MI) skills of a therapists* from the audio of a face-to-face counseling session. They present the results as an interactive visual dashboard that highlights strengths and weaknesses in the counsellors' communication. Finally, Wilbourne et al. [205] use ML tools to aid human coaches of a text chat-based app called Silby[4] to assess the quality and help improve their coaching response in real-time. However, the article does not report any system details or research findings.

In summary, the vast majority of papers described ML approaches to support the following: (i) the detection and diagnosis of mental health symptoms or conditions, (ii) predictions of mental health risks, or (iii) understanding of mental health-related behaviors (e.g., on online communities). Explorations of how ML could be leveraged outside of mental health assessment to support, e.g., (iv) mental health treatment or (v) health professionals remain scarce.

[4]https://www.sibly.co/.

3.2 Motivations for Applying ML to Mental Health

The following interconnected themes summarize motivations for applying ML to the domain of mental health.

3.2.1 Easy, Timely, Unobtrusive Access to more Objective, Scalable Mental Health Data. The use of social media [64, 165, 211], sensors [23, 27, 128, 168], and other technology interaction data [154, 222] has been described as allowing for the "non-burdensome," "unobtrusive," or "passive" assessment of peoples" mental health. These systems were suggested to enable "honest sharing of mental health concerns" (p. 754) [64] and to provide "natural data" as it is "generated by individuals in the normal course of their lives" (p. 10655) [133]. Sensor data was particularly valued for enabling the automatic, longer-term tracking of a person's mental health-related behaviors [44]. Social media content was claimed to present a "true reflection" (p. 358) [86] and "an unbiased collection of individuals' language usages and behaviours" (p. 1652) [33]. Further, such data was reported to be easy-to-access and retrieve; to offer a route to timely information for timely interventions; and to allow for data collection to be realized at scale [33, 165, 168, 222]. The analysis of data that is generated as part of peoples' everyday technology interactions and digital content creation was also reported to help identify objective markers [23, 168, 201] and systematic tools for capturing [61, 135, 184, 192] mental health behaviors, or assessing the skills of health professionals [78]. This argument was mostly justified through descriptions of the disadvantages of traditional questionnaires, interviews, self-report and survey tools with regards to: sampling biases, subjective reporting biases, risks of incomplete information, or underrepresentation [64, 78, 128, 153, 155, 211].

Despite much enthusiasm for easier, timely, and purportedly less biased data capture; one paper questioned the validity of developing diagnostic models for mental health conditions based on proxy data (e.g., a person's participation in a mental health community) rather than clinically validated diagnostic information [53].

3.2.2 Time and Cost Savings: Reducing Burden on Participant or Patient Effort and Clinician Time. ML approaches have also been described to potentially provide advantages in terms of saving time and costs. They can reduce efforts demanded of study participants or patients (e.g., to self-report) [176] and provide alternatives to clinical assessments [64, 128], which can be confined to health care professionals and specialized clinics; and thus, be expensive in terms of clinician time [23, 63, 122, 222]. In contrast, the collection of mental health proxy data from public social media is described as inexpensive, as it can be gathered with low effort and does not require any direct engagement with individuals [53].

3.2.3 Towards more "Accurate" and Reliable Mental Health Practices and Clinical Decision Making. Social media and sensor data has mostly been analyzed to help support, or speed up early detection, diagnosis, and treatment of peoples' mental health [6, 23, 31, 33, 86, 89, 193, 211]. However, for structured data and text analysis outside of social media, there was a stronger emphasis on the *need to advance existing health care practices by developing more "accurate," "reliable," and "evidence-based" clinical assessment tools.* For example, where text was analyzed to better understand or predict (acute) suicide risk, the need for novel, data-driven tools was argued by *foregrounding the insufficiency of existing clinical approaches*; suggesting that "traditional methodologies deployed in assessing suicide have not lived up to promise" (p.1) [2]. As a consequence, clinicians "are often left to manage suicidal patients by clinical judgment alone" (p.96) [145], and "are not able to reliably predict when someone is at greatest risk" (p.1) [134]. Similarly, papers that analyzed health records and questionnaire responses were often motivated to develop "automated," more "reliable," "less labor-intense," and "interpretable" diagnostic or risk assessment tools *needed to*

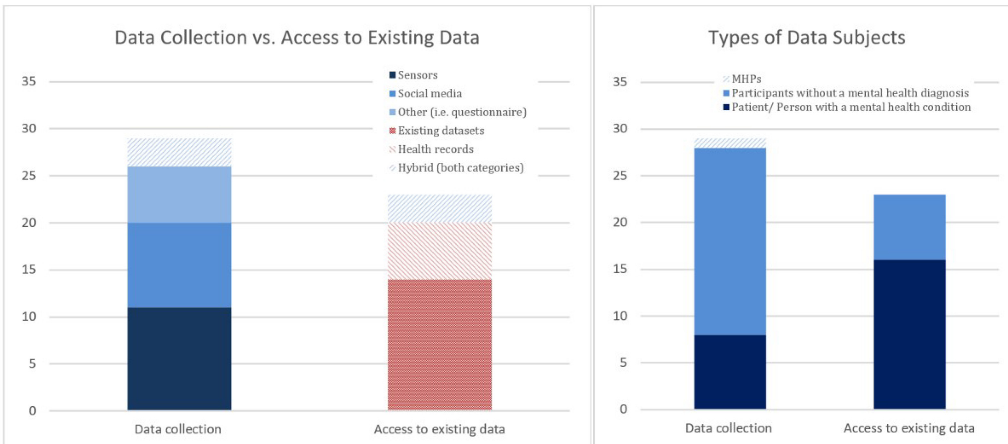


Fig. 4. Left: Proportion of papers reporting data collection or access to existing data. Right: Types of data subjects included in data collection or retrieved from existing data sources.

improve existing diagnostic or clinical decision making practices [61, 63, 112, 136, 139, 144, 179, 192]. For example, Tran et al. [192] made explicitly the point that, through their experiments in detecting suicide risk patterns from patient history, they demonstrated how their “proposed framework outperforms risk assessment instruments by medical practitioners” (p.1410).

In summary, key motivations for the use of ML for mental health include the following: (i) the possibilities afforded by access to behavioral data which is collected continuously and non-invasively; (ii) advantages of timely and automated data processing for efficiency and cost savings; as well as (iii) claims that data-driven assessments provide more objective, accurate and reliable assessments that can improve (clinical) practices and decision making. Thus, the literature often argues that novel models have advantages over existing approaches.

3.3 Data Scale, Subjects and Access in Mental Health

In this section, we outline how mental health data has been accessed or collected in the works reported; including details on the scale and from whom the data was sourced.

3.3.1 Source and Scale of Mental Health Data. ML algorithms build mathematical models based on training data to make predictions or decisions without being explicitly programmed [93]. The papers in our corpus are split between those that *collect data* for this purpose (n = 29) and those that *make use of existing data* (n = 23). Existing data is provided through previously generated datasets (n = 14, plus 1 hybrid) and health records (n = 6, plus 2 hybrids). In both categories, we identified three hybrid papers that described both a study of data collection and existing data use [45, 168, 222]. These have therefore been added to both category counts (see Figure 4). Very few records (n = 5) did not describe any data access or processing [28, 77, 82, 154, 205].

For the 29 records that collect data (see Figure 4, right), eight described recruitment or sampling of “patients” and “people with specific mental health conditions.” The remaining records primarily captured data from individuals who were described as “normal users,” “healthy subjects,” “students”, or “older adults” [44, 67, 139, 140, 153, 168, 179, 193, 201, 217, 218, 222], or for whom data was sampled from public social media (n = 8, plus 1 record that also includes a diagnostic sample [53]). One record further collected audio data from mental health professionals (MHPs) [78]. Table 2 provides an overview of the numbers of people, including those included as

Table 2. Overview of the Number Individuals Who Were Included in the Respective Studies

Data Collection			Data Access	
Patients/People with a Mental Health Condition	Participants without a Clinical Mental Health Diagnosis	MHPs	Patients/People with a Mental Health Condition	Participants without a Clinical Mental Health Diagnosis
10 Patients with bipolar condition [50] 24 Patients with PTSD [23] 25 People with MDD [31] 26 Students with suicidal history [134] 59 Individuals (31 people with schizophrenia, 16 people with mania, 9 people as control group) [45]—Study 1 66 Notes (33 notes of people who committed suicide + 33 notes of people who simulated a suicide note) [145] 90 Patients in a psychiatric hospital [4] 143 Individuals (88 patients with schizophrenia + 55 individuals as control group) [53]—Study dataset 4	Use of clinical screening tools: 7 Older adults [153] 20 Participants (reported interest in stress management) [**140**] 26 Healthy adults [67] 33 Students [44] 48 Students [201] 105 Students [168]—Study 1 108 Students [217, 218] No use of clinical screening tools: 7 Participants (under-defined) [193] 10 Participants (high scores on stress scale) [139] 27 Participants [222] - Study 1 200 Twitter users (100 mental disorder + 100 random) [86] 585 Twitter users with self-reported diagnosis of depression [33] 656 Participants (300 students + 353 working professionals) [179] 1,965 Twitter users (1,426 who self-report schizophrenia + 539 individuals as control group) [53]—Study datasets 1–3 4,000 Twitter users (2,000 who self-report symptoms of depression + 2,000 people as control group) [**211**] No numbers of research individuals reported for [57, 89, 133, 141, 165]	21 Counsellors [78]	5 Patients with symptoms of depression [222]—Study 2 20 Individuals (12 people with bipolar condition, 8 people as control group) [27] 55 Individuals (23 people with unipolar or bipolar depression, 32 people without symptoms of depression) [62] 66 Participants of EASE dataset [112] 79 Psychiatry reports [45]—Study 2 84 Patients [122] 91 Patient records [63] 130 Patients [2] 142 Individuals from DAIC-WOZ database [168]—Study 2 196 Outpatient mental health records [144] 197 Patients with bipolar condition [61] 201 Patients from various reference datasets [6] 275 Individuals from E-DAIC dataset [155] 1,090 Hospital patients with symptoms of depression and comorbid conditions [136] 7,746 Patient EMRs [192] No numbers of research individuals reported for [184]	Use of clinical screening tools: 805 Participants (48 students, 757 information workers) [128] 16,952 Older adults (2191 with and 14751 without symptoms of depression) [209] No use of clinical screening tools: 224 College students [135] 566 Mobile phone users [177] 7,261 Users of a commercial wellness platform [152] 17,251 Mobile phone users [176] No numbers of research individuals reported for [64]

"control" groups, that were studied in each data experiment. Due to their data sampling approach (described below), many of the social media papers did not specify any "user" numbers; and instead report the "total number of posts and comments" that were analyzed (e.g., 3000 [89], 4026 [57], 5000 [133], 7410 [141], 113,337 [165]). Contrary to this, data that is accessed as part of existing datasets and health records predominantly included information about "patients," or "people with a mental health condition" (n = 18); and to a lesser extent individuals without a clinical mental health diagnosis (n = 5) such as: mobile phone users [176, 177], students and workers [128, 135], social media users [64].

Table 2 further outlines the number of people that were included in the respective studies. It shows that, outside of the analysis of health records (e.g., [136, 192]) and specific existing large-scale datasets [152, 209], the number of "patients" or "people with a mental condition" included was generally quite low, especially when considering that advanced ML approaches require a lot of data. Next, we outline how the papers approached the collection and access to data, and how they conceptualized it as mental health data.

3.3.2 Data Collection and Related Conceptualizations of Mental Health. Only a small proportion of the papers (n = 8) recruited either *psychiatric hospital patients* [4], or people with a *diagnosed* *mental health condition* [23, 31, 45, 50, 53, 134, 145]. For example, Pestian et al. [145] described the process by which three MHPs conducted linguistic annotations of notes written by people who "committed suicide" and compared those to people who "simulated" writing a suicide note as "controls." The authors however do not mention how they obtained access to real suicide notes. In contrast, Nobles et al. [134] actively recruited individuals with a history of past suicidal thoughts and behaviors. In a lab study, participants downloaded and labelled all outgoing SMS messages to identify events of attempted suicide, suicidal ideation, or depression. Psycholinguistic features and word occurrences in the SMS texts were then analyzed to identify cues of heightened suicide risks.

A significant proportion of the papers (n = 21) described data capture studies that involved people who may not have a mental health problem, or diagnosable mental illness. Thus, to define and extract mental health specific behaviors (e.g., from sensor and interaction data) a number of approaches were applied. Most commonly (n = 8), the researchers used (i) *questionnaires or standardized clinical scales*[5] to screen for specific mental health symptoms and their severity within a study population [44, 67, 140, 153, 168, 201, 217, 218]. For assessments of symptoms of depression, this commonly included the CES-D [153], BDI [44], and PHQ [140, 201, 217, 218]. For symptoms of anxiety, reported instruments encompassed the STAI-Y [67], SIAS, and SPS [168]; and for symptoms of stress the PSS [201, 217, 218]. In a few instances, the researchers further employed (ii) *experimental scenarios* to induce and control for specific experiences in study participants such as stress [67], anxiety [168] and emotional states [222]. For example, Salekin et al. [168] approached their data collection by using various scales to assess the social anxiety of university students about public speaking, and divided them into a low and high anxiety group. Later, participants had to quickly prepare a 3-minute speech and present it in front of a large video camera. Audio of their speech was then analyzed to detect "socially anxious speakers." In addition, (iii) *ecological momentary assessments*[6] *(EMA)* were regularly applied to evaluate users' experiences and support data labelling [44, 67, 128, 139, 140, 176, 218].

Of the thirteen papers that described data studies involving individuals for whom *no clinical screening tools were used to assess their mental health status*, nine presented an analysis of social media content. These works extracted data from public posts, mostly using specific Reddit or Twitter APIs [33, 89, 141]. Only in one instance, there was direct engagement with social media users to recruit individuals with clinically assessed schizophrenia from inpatient and outpatient psychiatric departments [53]. For mental health diagnosis or the detection of specific mental health states, these works primarily prospected for different types of "diagnostic signals" in online social behaviors that can be grouped into the following: (i) *affiliation behaviors*, (ii) *self-report*, and (iii) *external validation* (see framework by [53]). Here, most papers (n = 6) focused on affiliation behaviors, whereby membership to an online mental health community [57, 64, 89, 141, 133, 165], or engagement with mental health content (e.g., using hashtags of #anxiety, #depression

[5]Examples of assessment scales used include, for *depression*: PHQ-8 and PHQ-9 [97], Epidemiological Study Depression Scale [9], and MADRS depression rating system [126]; for mania the Mania Rating Scale (MRS) [213]; for *mood* HAMD mood scores (Hamilton rating scale for depression) [50]; for *affect*: PANAS for positive and negative affect [202], Photographic Affect meter (PAM) [148]; for *stress*: Coping Stress Questionnaire (CSQ) [159]; Trauma-Focused Coping Self-Efficacy (CSE-T) [16]; Perceived Stress Scale (PSS) [38]; PTSD severity checklist (PCL) [203]; for *mental wellbeing*: eight-item flourishing scale [46] and SF-36 Mental Health Score (www.optum.com/sf36); for levels of *social inclusion and connectedness:* 20-item UCLA loneliness scale [163]; and for *physical activity*: Yale Physical Activity Survey (YPAS) [47].

[6]EMAs are often short questions designed to capture in-situ real time information about a person's experience [128]. Examples of EMA's used include the following: Experience Sampling Method (ESM) based on two-dimensional Circumplex model of emotion [162]; PAM picture library to assess mood [151]; EMAs built on single item stress survey [182]; Stress Monitoring Test (SMT) [139].

or #stress [86]) are treated as a *proxy* for diagnostic information. The remaining papers (n = 3) identified users with a diagnosis of depression through public self-report of a mental illness diagnosis [33, 211]; for example by pooling Twitter posts of people who stated "I was/have been diagnosed with depression" [33]. Across these examples, we found no evidence of external validation of assessed diagnostic signals through, e.g., clinical appraisals or clinical scales, with exception of Ernala et al. [53]. The authors [ibid] contribute an empirical study that assesses and compares the internal and external predictive validity of different social media-derived proxy diagnostic signals for mental illness diagnosis of schizophrenia (see further Section 3.5.1). In other works, expert assessment was used to help validate proxy signals, or guide data analysis [64, 89, 165, 211]. For example, Yazdavar et al. [211] developed with psychologists a lexicon with 1,620 depression-related symptoms (categorized based on clinical PHQ-9 symptoms of depression [97]) to guide their analysis.

Finally, Hirsch et al. [78] described automatically extracting insights about the MI skills of counsellors from audio signals. Initially, session recordings were labelled using an established MI Skills Code (MISC) [121]. This was then combined with speech signal processing to generate an *MI quality* score (composed of measures of: empathy, MI spirit, reflection-to-question ratio, and others—as informed by the MI Treatment Integrity Scale [129]).

Thus, across all data collection papers, we found a range of different approaches for capturing, processing, and labelling data to help isolate indicators of mental health, or facets in the communication skills of MHPs. Although many papers targeted detection and diagnosis of mental health conditions, outside of recruiting patient populations and explicitly applying clinical measures, there was rarely any (external) diagnostic validation of the assessed phenomena.

3.3.3 Access to Pre-existing Mental Health Data as an Alternative to Data Collection. Fifteen papers reported utilizing pre-existing datasets to train predictive models or develop new ML approaches [6, 27, 62, 64, 112, 122, 128, 135, 152, 155, 168, 176, 177, 184, 209]. This included the use of various resources of multi-modal data such as the Analysis Interview Corpus (DAIC-WOZ) [196, 184] and its extended E-DAIC dataset [155]. These datasets contain audio-video recordings of clinical or AI agent-conducted interviews with people who experience psychological distress conditions, such as anxiety, depression, and PTSD. Other examples include the AVEC 2013 audio-visual dataset for studies of depression [122, 168], and the EASE dataset of people undergoing trauma treatment, e.g., for PTSD [112]. The BiAffect mobile phone and Depresjon dataset were used to access acceleration data of people with depression [62] and bipolar conditions [27], while the English Longitudinal Study of Ageing (ELSA) provided psychological and mental health data on older adults as indicators of depression [209]. Finally, a few papers reported on the re-use of previously collected user data in the context of a commercial wellness platform [152], for social media analysis [64], and mood or well-being research (e.g., Emotion Sense [176, 177], SNAPSHOT[7] [135], and StudentLife[8] [128]).

Alongside existing datasets, a number of papers (n = 8) accessed (electronic) health records and other clinical notes, recordings, or reports for their analysis. These records can provide an important data resource as they can document a wealth of information about demographics and care delivery such as admission dates, types and frequency of interventions, and the results of clinical assessments. However, the papers provided few, if any details on how access to health record data was negotiated. Zhou et al. [222] for example only mentioned having been provided

[7] https://snapshot.media.mit.edu/info/.
[8] https://studentlife.cs.dartmouth.edu/.

with video and audio chat content of patients with symptoms of depression by a psychiatrist; while other papers [136, 144, 192] only stated the type of hospital, health department or services from which data was received. Often, patient information was solicited from a hospital [2] or other mental health service, institute or psychiatry department [45, 61, 63] that at least one or more of the paper authors were affiliated with. This further suggests that this kind of data access and analysis may be primarily led and conducted by health organizations or requires close collaboration with these institutions and care providers.

3.4 Types of ML Techniques Used and Model Evaluation Approaches

Next, we outline the ML techniques used in the papers, and how generated ML models were evaluated.

3.4.1 Machine Learning Tasks and Techniques: Primarily Classification and Supervised Learning. A number of different ML techniques can such as classification, regression, association, and clustering can be applied to common tasks such as identifying correlations and pattern recognition in high-dimensional datasets to achieve more simplified, human-interpretable formats [22, 136]. Building on the approach by Shatte et al. [173], we grouped the papers in our corpus into four ML-algorithm categories: (i) *supervised*, (ii) *unsupervised*, (iii) *semi-supervised learning*; and (v) *novel techniques* (see Appendix A1 for an overview).

The vast majority of the papers in our corpus (n = 37) used supervised learning, and most often described the application of one or more of these techniques: Support Vector Machines, Random Forest, Decision Trees, *k*-Nearest neighbors, supervised LDA, Lasso, and Logistic Regression [23, 33, 44, 45, 50, 53, 57, 61, 62, 67, 89, 112, 122, 128, 133, 135, 139, 145, 155, 165, 176, 179, 184, 192, 193, 209, 218, 222]. For supervised learning, data is labelled and then used to train a model that then can predict the label for new data. Here, the dataset contains both the inputs and desired outputs. In our corpus, supervised learning was primarily applied for *classification* tasks, whereby a set of previously classified training instances is used to build a model that can predict, for example, a binary class label (e.g., presence or absence of a symptom), or a limited set of class labels (e.g., mental health condition) of unseen instances.

Unsupervised learning uses mathematical techniques to *cluster* data to provide new insights. Here, the dataset only contains inputs, but no desired output labels. To discover patterns and help structure the data, clustering methods respond to the presence or absence of commonalities in each piece of data. Across all papers, only two specifically applied clustering to distinguish language use in online communications [141], and extract diagnostics from psychiatric reports [45]. However, most often, data clustering was applied as an initial step to classification (described above) to aid the selection of features or identify labels for developing supervised learning classifiers [2, 63, 64, 89, 122, 139, 155, 176, 179].

Only one of the papers explicitly described the use of semi-supervised learning techniques [211] that combine labelled and unlabeled data in their model; and few papers (n = 6) reported the use of novel methods. Novel methods included the application of custom ML models to create multi-dimensional classifiers [136, 152] or to forecast mental well-being from sparse self-report data [177]; deep learning (DL) methods [86, 134]; and reinforcement learning (RL) to help create personalized recommendations for a stress-management interventions [140]. The remaining papers either described proposals or concepts that did not apply any ML [28, 31, 77, 82]; or applied existing classifiers to newly collected data [201].

Finally, the *analysis of natural language (NLP), speech and text,* presents a specialized area of ML that mostly utilizes unsupervised techniques. Various works applied lexicon- and other text-mining approaches (e.g., [71]) to help extract keywords (i.e., depression), topics, or linguistic

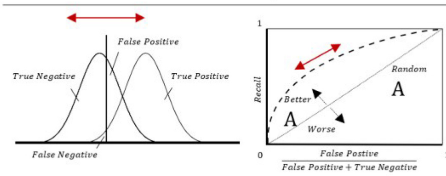
$Accuracy = \frac{True\ Postive + True\ Negative}{Total\ (Positive + Negative)}$	*Accuracy* is the measure of correct classified test samples against the total number of test samples. Averaged for each of the classes, it provides a measure of the accuracy of the entire classifier.
$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$	*Precision* is the classifiers ability to not label a sample as positive, if it should be negative. It is calculated as the ratio between true positives, divided by the sum of true positive and false positives.
$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$	*Recall* represents the percentage of positive samples that were correctly labeled as positives.
$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$	*F1 score* is calculated as a weighted harmonic mean of the classifiers' precision and recall (F1 = 1 is best).
	*AUC (Area under the ROC curve)* is the probability for a random positive and negative example, that the classifier will assign a higher score to the positive example than to the negative. It measures the entire two-dimensional area underneath the ROC curve. The larger the AUC area the better (AUC = 1 is best).

Fig. 5. Definitions of the five most commonly used performance measures for classification accuracy/error.

features from text to learn high quality features from human speech or text to develop different classification models, or determine its semantic polarity [2, 33, 45, 64, 89, 133, 134, 141, 145, 155, 165, 211]. A small number of works (e.g., [23, 78, 122]) also analyzed acoustic, para-linguistic features in speech such as estimates of prosody, pitch, or speech rate.

Thus, in keeping with the majority of the papers' focus on mental health assessment, the works primarily applied supervised ML techniques to investigate if, and how well, certain mental health behaviors, states, or conditions could be classified through newly developed data models. Most unsupervised learning techniques were applied to support data labelling and feature selection for classification. This is in keeping with clinical systematic review findings by Lee et al. [100] and Shatte et al. [173]. Other routes to leveraging ML techniques e.g., for enabling personalization, however, remain under-explored.

3.4.2 Performance Evaluation of ML Models: Common Techniques and Performance Measures. As described, labelled classification tasks were most prominent in our corpus. The performance of developed classification models is typically evaluated by their ability to generalize classifications or predictions to new cases, meaning how accurately a classifier predicts the correct class labels for new data, for which the desired output is known. To achieve this, part of the available training data is typically "held-back" (not included in training the model), and instead used to test how well the model performs on that held-out data (cf. [149]). For this, the papers reported evaluation techniques of *Leave-one-user-out (LOSO)* or *k-fold cross-validation* [2, 4, 6, 23, 27, 33, 44, 50, 53, 57, 61, 62, 63, 64, 67, 78, 86, 89, 122, 128, 133, 134, 136, 139, 140, 144, 145, 152, 153, 165, 168, 176, 177, 179, 192, 209, 211, 218, 222].

To report the performance of developed classifiers, the majority of papers reported measures of *accuracy, precision, recall,* and *F1-scores* [4, 6, 27, 33, 45, 53 57, 61, 64, 89, 135, 134, 153, 165, 168, 179, 192, 222]. In a few instances *Log Loss* was used that considers the uncertainty of a prediction based on how much it varies from the actual label [209]. The measure of precision indicates how useful a prediction is (low false positive rate) and recall how complete it is (low false negative rate). Accuracy is the measure of how many samples or individuals are correctly classified out of the total number classified, and the F1 score is a calculated weighted harmonic mean of the classifiers precision and recall (Figure 5). For imbalanced datasets with unequal error costs, the *area under the ROC curve (AUC)* metric was often used (cf. [53, 61, 139, 152, 176, 218]) and described as a more appropriate evaluation technique. In a few instances of regression tasks[9] [27, 112, 122, 128, 155, 193], metrics of *mean error* (e.g., MSE, MAE, RMSE, SMAPE) were applied [50, 62, 112, 155,179] to reveal any unexpected values, sensitivities towards outliers, and risks of over- or under-estimating false predictions [51]. Individual works also applied more specific metrics to evaluate multi-dimensional classification (i.e., using *Hamming score, Hamming Loss, Exact-match* [136]); or the confidence [139], coherence [64], and completeness [45] of clustering outcomes (e.g., using *WCSS, Dunn Index, DB Index, or Silhouette Index* to assess similarity within, and separation between, clusters [179]).

In summary, developed ML models were commonly evaluated using aggregate metrics such as accuracy, AUC, and mean square error. While these metrics present established performance measures, as aggregate measures, they can hide varying model performance or biases across different population groups (cf. [73]). This also emphasizes the need to ensure that existing datasets capture the complexity of the real world (e.g., to not under-represent certain groups); especially given that the papers in our corpus generally assess the generalizability of achieved models by using parts of their training data.

3.5 Research Insights and ML Specific Challenges

Next, we summarize the three main types of contributions that were reported from the research and developed models and provide a brief overview of commonly reported ML-specific challenges.

3.5.1 Research Insights. Although the majority of the papers described motivations to help detect or diagnose mental health problems to impact health management practices (see Section 3.1), the vast majority of the papers (n = 42) primarily focused on the *technical or algorithmic development of (initial) ML models.* Here, success of newly developed models is primarily reported through performance measures of the “accuracy of the (best) classifier” or “robustness of clustering” [2, 4, 6, 23, 27, 33, 44, 45, 50, 57, 61, 62, 63, 64, 67, 86, 89, 112, 122, 128, 133, 134, 135, 136, 139, 141, 144, 145, 152, 153, 155, 165, 168, 176, 177, 179, 184, 192, 209, 211, 218, 222]. To further demonstrate how newly developed models “outperformed existing ones” (with few exceptions [44, 135, 136]), the performance metrics are often compared with default or baseline models, and other state-of-the art approaches [2, 27, 33, 62, 122, 134, 135, 152, 165, 168, 176, 177, 211, 218]. In addition to performance reports, a number of these papers foregrounded methodological contributions such as: new approaches to data labelling [168, 211] and feature extraction [112]; the inclusion of time-dependent features [33, 165, 192, 211, 218]; improvements to data representations [177] and data integration [27]; and strategies to optimize data collection (periods) [152, 128, 218]. Building on these results, authors often concluded how their *work presented a “proof-of-concept” that showed “the potential” of using a particular technology* [27, 50, 153], *data source* [57, 62, 64, 112, 128, 134, 165, 176, 218], *or algorithm* [44] *for understanding, detecting or inferring a relationship with mental health.*

As a second contribution type, a number of studies sought to advance our *understanding of mental health.* To this end, they extracted the “importance” of identified features and their “relations” with mental health [50, 63, 89, 135]; or complemented their ML analysis through the reporting of qualitative findings [139, 133], visualizations [57, 141], and other user information [201] to contextualize and aid the interpretation of ML outputs in relation to mental health. Especially for understanding online mental health communications, the works illustrated how the identification

[9] Like classification, regression is a predictive modelling task. Although classification predicts a class label for a given observation, regression instead predicts a continuous quantity (e.g., amount, sizes, ranges, time series).

of discussion topics can inform the design of online mental health interventions [64, 141], assist moderators of online communities in prioritizing their responses [28, 89]; and informing education and intervention strategies [165, 133, 141]. Outside of online social media, Yang and Bath [209] for example calculated what features derived from questionnaire data were particularly related to symptoms of depression in older age. Of the top 80 identified influential features, they found nine key factors, including "loneliness" and "quality of life." Aiming to advance our understanding beyond individual factors that can impact peoples' behaviors or mental health, Nosakhare and Picard [135] studied *what combinations of health behaviors may lead to certain health outcomes*. To this end, they analyzed stress experience patterns from multi-modal data and extracted "behavior combinations" that had the highest probability of co-occurring. Although innovative in approach, the *insights achieved by these works were often preliminary and require further research to substantiate.*

Thirdly, of all 54 papers that we analyzed, few reported empirical research findings of ML models or deployed ML interventions [53, 78, 140]. This includes work by Ernala et al. [53] who assessed the clinical validity of ML models that were developed based on "proxy" diagnostic information sourced from social media. The authors found that the predictive models that were based on this proxy data had strong internal validity but performed poorly when tested on the social media data of people who had a clinically diagnosis (poor external validity). ML models built on affiliation behaviors alone (e.g., being a follower of a Twitter account that focused on schizophrenia) were reported to have the poorest performance. Their study also revealed that the inclusion of clinical judgment to appraise self-reported mental illness on social media showed the best performance among the three tested proxy signals [ibid]. This work therefore contributes to important discourse about *construct validity of captured data* and the *importance of involving clinical expertise and assessments* for developing accurate and reliable ML-supported diagnosis.

Furthermore, Parades et al. [140] conducted a 4-week study of a mobile app to explore how ML could be utilized to personalize a stress-management intervention. Their experimental study design varied app recommendations to be either driven by the "ML" or "randomly" selected; and whether the user "can" or "cannot" self-select the recommended intervention among other options. The results showed how both ML conditions had the greatest and statistically significant positive impact on stress reduction. Yet, their findings also showed how the ML algorithm reduced the diversity of the intervention recommendations over time. To avoid boredom and attrition; the authors suggest "adding diversity" as an objective to the ML algorithm.

Finally, Hirsch et al. [78] reported the findings of a study with 21 counsellors evaluating an interactive user interface that visualizes the output of a system that automatically assesses their MI skills from audio. They evaluated how counsellors responded to the concept of automated skills assessment, how the system may fit within or disrupt their clinical practice, and what concerns they may have. Results indicated difficulties for counsellors to understand some of the global measures (i.e., how "MI spirit" was derived from the data); as well as perceptions of system-derived data as being "objective" and "hard to contest." More experienced counsellors were also more likely to question the accuracy and calculation of system feedback; and there was a desire for "actionable" feedback to help improve their skills.

Despite these works reflecting single instances in our corpus, they help initiate important discussions of the role, acceptance, and broader implications of positioning ML systems within peoples' work practices, and lives. They also begin to show how deploying and studying ML systems in real-world mental health care contexts is needed to inform and test the design of useful and effective ML-enabled interventions.

3.5.2 Frequently Reported Data and Modeling Challenges. One of the most frequently described data challenges has been the *capture of accurate, reliable mental health data* (n = 22) due to "noise, ambiguity, or redundancies" in the data [28, 31, 112, 136, 153, 165, 168, 176, 177, 211, 222]; and difficulties to identify "robust labels" for "subjective, non-discrete human experiences" [64, 67, 89, 135, 153, 155, 168]. This challenge was particularly pronounced where information to help (clinically) validate assessed phenomena was missing [31, 53, 133, 165, 168], and also for research aimed at identifying data models that are transferable to other (real-world) data contexts [23, 45, 67, 218]. In terms of ambiguity in data, Rabbi et al. [153] and Hirsch et al. [78] describe difficulties to isolate, in audio signals, the speaker of interest from the environment (e.g., speech emitted from a television). For text analysis, ambiguous terms like "depression" were described as a challenge as it also describes the "economy," a "historic era," and is semantically difficult to separate from expressions of transient sadness: "I am depressed, I have an exam tomorrow" [211]. Informal language (e.g., word repeats "yayayay"), abbreviations (e.g., "ikr" for "I know right"), colloquialisms and improper sentence structure (e.g., "hehe thanks") [134] further add complexity; alongside elaborate lexical variations that people deliberately develop to undermine communication bans online (i.e., changing "thighgap" to "thyghgappp" in eating disorder contexts [28]).

A large proportion of the papers (n = 23) also *acknowledged varied limitations of the dataset(s) they worked with*, primarily suggesting a need for "larger datasets" (e.g., [61, 82, 89, 128, 145, 155, 177]) to compensate for missing or sparse data, and to be mindful of noise and errors in data recording. Also acknowledged was a limited generalizability of the established results due to restrictions in the study sample [2, 45, 61, 89, 134, 136, 141, 152, 211, 168, 222] and uncertainty about other unknown confounding variables [57, 165, 168]. For example, Yazdavar et al. [211] acknowledged that their focus on social media data meant they would only capture people who generated ample content online and were open to expressing symptoms of depression publicly. In addition, a number of papers mentioned concerns about biased, missing, or incomplete data [6, 45, 45, 53, 78, 128, 141, 176, 177]. Risks of potential biases were most elaborated by Ernala et al. [53]. The authors conducted an error analysis that revealed how statistical data distributions can be drastically different between the social media proxy datasets that they analyzed, and actual patient datasets, which foregrounded "population and sampling biases." Additional linguistic analysis also showed how patients with a clinical diagnosis of schizophrenia, in contrast to the Twitter users studied, largely had private Facebook accounts and did not exhibit disclosures about their schizophrenia experiences and support seeking behaviors on their social media. This brings into question to what extent proxy diagnostic data can, indeed, provide clinically grounded "diagnosis" information about a person [ibid].

Furthermore, outside of data processing challenges specific to the respective ML tasks and techniques applied in each work, a number of records (n = 8) explicitly outlined *difficulties with integrating varied, often multi-modal data sources* [27, 122, 128, 139, 154, 155, 192, 222]. For example, Tran et al. [192] described the complexity of working with temporal medical databases that host multiple time-indexed records for each patient that can include: sparse and irregular episode data; heterogeneity in patient records; distribution shifts (i.e., new record keeping or treatment procedures), and many other types of information (i.e., demographics).

Finally, some of the records (n = 9) *acknowledged limitations with regards to the ML modeling approach that was chosen* by advising caution regarding the use of retrospective data for predicting future behavior [144], acknowledging that current classifiers were designed to detect presence, duration or frequency of symptoms, but not symptom severity [211], and proposing the use of more "personalized approaches and individualized models" to more accurately assess experiences of mood [50] or stress [67], and support efficiency in detecting specific mental health conditions [133, 152]. Some works [165, 201] were mindful in their reports about *difficulties with speculation about the cause and effect* of achieved, often correlation-based results that do not permit any causal claims [128]; whereas others described the potential *implications of errors* in model predictions [50, 53, 78, 134, 218] (see further Section 3.6.4), or mentioned *needs for data security* through secure data storage and handling [6, 82, 153, 201]. See Appendix A2 for further detail.

In summary, the works described a number of common data modeling challenges. Primarily, these included the following: (i) difficulties to robustly measure and label peoples' mental health as a complex, multi-faceted, and dynamic concept from often noisy or ambiguous data, (ii) technical challenges in generating low-dimensional features that reduce (initially perhaps richer, diverse) data sources into a small number of quantifiable categories suitable for modeling, (iii) choices in model selection and training algorithms, (iv) acknowledgement of needs for "more data" to increase model accuracy and generalizability, and (v) to reduce risks of errors or biases.

3.6 Ethical and Research Issues in Real-World Applications of ML

This final section describes the extent to which developed ML models (i) were envisioned for, and used within real-world mental health contexts, (ii) followed user-centered methods in their design or study, and (iii) described any design challenges and ethical issues regarding the research or deployment of ML-systems.

3.6.1 Real-World Use (and Potential Implications) of Developed ML-Models or Applications. As described previously, only two records described user studies of ML-applications [78, 140]. Despite few examples of ML-enabled systems in-use, a substantial amount of the records included *speculative descriptions, proposals, and claims* how developed ML models may come to impact on clinical or everyday health management practices [6, 27, 33, 64, 89, 112, 128, 134, 136, 141, 152, 165, 168, 176, 177, 179, 211, 209, 218, 222]. For example, Zhou et al. [222], who suggest the development of multi-modal sensor systems to unobtrusively assess mental health from everyday technology interactions, described the potential impact of their work as follows: *"We expect that the outcome of this research will be an effective tool for assessing the affective states of individuals on a large scale. It can be used as an enabling component for developing new mental health solutions, including identifying the onset and severity of mental health problems in individuals and may prove to be of use to clinicians, for self-awareness, and for support from family and friends"* (p. 1402). Similarly, Salekin et al. [168] suggested: "*The ability to identify symptomatic individuals from their audio data represents an objective indicator of symptom severity that can complement health-care providers' other assessment modalities and inform treatment";* and claimed their *"framework is a scalable complement to health-care providers' self-report, interview, and other assessment modalities"* (p. 21f).

Thus, despite strong motivation for developing ML approaches that can make a real difference in this important domain, only a very small number of works sought to introduce developed data tools and insights into real-world settings. This means that the actual impact of achieved ML models in terms of effectiveness and relevance for mental health; or use and acceptance by laypeople, remain—so far—mostly speculative.

3.6.2 Multi-disciplinary Research Teams and Engagement with User-Centered Design. In keeping with the review findings by Shatte et al. [173], we found that the majority of the papers were authored by multi-disciplinary teams (n = 29). This included experts from health and social sciences (i.e., medicine, psychology, psychiatry, behavioral, and educational sciences), engineering (i.e., computing science, data science, intelligent systems), and occasionally, arts and design [77, 78, 140]. Of the remaining works, a substantial proportion was authored by experts in computing (n = 22), and in few instances by experts in health (n = 1), psychology (n = 1), or social work (n = 1). Despite a *predominantly multi-disciplinary set-up within the research teams*, there was however *little reporting of user-centered design processes*. Notably, the work of Hirsch et al. [77, 78] presents the strongest example of research that followed both a participatory, iterative design process and presented a pilot study to evaluate their ML systems with prospective users. Mostly, user involvement was only sought in the collection of real-world user data (e.g., [4, 50, 136]), and discussed in terms of pragmatic challenges (e.g., requirements of keeping technology charged and used; and users' compliant with data collection [44, 152, 153, 176, 201]; software compatibility issues in data extraction from varied devices [44, 134]; and other technology infrastructure challenges [218]). Only a few studies described the active involvement of target-users, MHPs, or other domain experts in data labelling (e.g., [64, 134, 136, 211]) and for validating ML model results [53, 218]. For example, Chang et al. [31] adapted contextual inquiry [17] as a method to capture tacit diagnostic knowledge of MHPs in categorizing voice utterances of people suffering from major depressive disorder (MDD). Zakaria et al. [218] conducted semi-structured interviews and collected survey data during their data collection study to "verify" primary causes of student stress. This includes information about how the students were managing their stress experiences, and insights about their work meeting dates, duration, and location. Using such data primarily as "ground truth" to validate their models, there is limited reporting of the interview findings in the article.

The general restriction of user involvement to data collection and labelling suggests a gap in user-centric dialogue and more collaborative involvement with MHPs and people with lived experiences of mental health (problems) that could support a deeper engagement with important mental health needs, and also aid with the challenges involved in appropriately addressing these needs through ML-enabled insights or applications.

3.6.3 Designing Interpretable and Trustworthy ML Models and Applications for Mental Health. A key challenge for the use of ML-enabled outputs and systems within real-world mental health context is to ensure that non-ML experts are able evaluate the performance of ML models and decide whether to trust their outputs. However, only a few papers mentioned the need for future work to develop front-end interfaces for MHPs to present and interact with the ML outputs [192, 141], support clinician understanding of how certain data features influence model decisions [155], to "explain the reasoning" behind ML predictions [192, 177], and study the acceptance of proposed data tools by MHPs [82]. Spathis et al. [177] for example explain: "*Although the scope of model interpretability is very wide, including causality, informativeness, and transparency, at least post-hoc interpretations and visualizations are needed to qualitatively evaluate what a model has learned. This is especially relevant in clinical setups where clinicians can only rely on interpretable models to make informed decisions".* Furthermore, Tran et al. [192] described *"transparency in modeling decisions and interpretability in results"* (p. 1411) as a key modeling consideration and presented earning trust from clinicians for deploying their modeling solution in their daily work-flow as the main challenge: *"We anticipate that the initial resistance will be significant as the implication of taking the advice from the machine is profound for professionals"* (p. 1417).

Only one research project [77, 78] explicitly engaged with the design challenges of creating an interactive interface for presenting model outputs that are human interpretable. Here, user evaluation findings showed how more experienced MHPs were more likely to question the accuracy and calculation of system feedback and expressed a desire to be able to inspect and potentially dispute ML outputs that seem unreasonable. Further, there had been a tendency, especially of trainee MHPs, to uncritically accept system generated outputs as "objective measures," even when trainees acknowledged that they did not fully understand how feedback was derived, or what it precisely meant. Here, their willingness to trust the ML system was bound up with the perceived "legibility" of the output rather than its statistical accuracy [78]. Thus, the authors concluded that designers, especially when developing systems that can have potentially adverse impact on human welfare, carry the responsibility to create mechanisms that enable users to contest system outputs and suggest developing reasonably accurate models first, before deploying them in a health context.

Thus, outside of understanding and addressing ML model development challenges (Section 3.5.2), there is a need for more study of how model outputs are interacted with and become interpreted by laypeople—who may become the end-users or beneficiaries of ML-enabled solutions. Existing works further emphasize the importance for interface design to support an *appropriate* level of understanding and trust in the models.

3.6.4 Considerations of Ethics. Our final theme captures the extent to which the papers described or addressed ethical issues or procedures in their research. Echoing recent reports by Sanches et al. [170], we found that a significant proportion of the papers (n = 26) did not include any mention of ethics despite their focus on a sensitive area of health care research (Table 3). Of the remaining papers, a significant proportion (n = 15) primarily reported approvals or exemption form ethical review processes. Next, we expand on additional ethical considerations that were communicated, and how they relate to core ethical health care principles of (i) autonomy, (ii) beneficence and non-maleficence, and (iii) justice.

3.6.4.1 Autonomy (Including User Consent and Human Agency in ML-Informed Decision-Making Processes) A large amount of the papers addressed the value of autonomy through the *application of privacy protecting measures to respect, and ensure confidential treatment of, peoples' personal information* (n = 11) [27, 122, 128, 141, 153, 165, 168, 201, 211, 218, 222]. For example, with sensor-based data capture, authors often chose to only record or process higher level data such as the number or duration of specific phone interactions [27, 128], or audio features from human speech [122, 153, 201] rather than any typed or spoken words to preserve users' privacy. Here, Rabbi et al.'s [153] described how such measures not only enabled data capture in a realistic user environment but were also perceived as user-friendly: *"Although the recorded features do not allow reconstruction of audio afterwards, they enabled us to infer when human voice was present and whether there was conversation. (...) it is worth mentioning that during the study we learned that the privacy sensitive audio data collection was very well accepted by users"* (p. 387).

Similarly, for social media data, some of the authors acknowledged the analysis of potentially sensitive behavioral health information. They justify their data use by reporting to have pooled "publicly" available, so called observational data, whose data collection did not involve any interaction or intervention with subjects [141, 165, 211]. The argument is thus that such usage does not require explicit user consent. For example, Saha and De Choudhury [165] described how no direct contact was made with users who posted in the subreddits they analyzed and that it was deemed impractical to gain informed consent from thousands of people. The authors acknowledged that *"therefore individuals may be unaware of the implications of social media content, with regards to their ability to signal underlying psychological risk"* (p. 23). Outside of social media studies, proposals for the need for users to take control over their data use for diagnostic assessments were rare [82]. Few papers explicitly mentioned user consent processes for primary data collection (n = 5) [23, 61, 179, 209, 222]. Among those that did not explicitly mention consent were studies that reported patient interviews in a psychiatric hospital [4]; the analysis of mental health records [45]; or audio and video recorded conversations between patients with symptoms of depression and their psychiatrists [222]. Here, arguably, requirements for consent are balanced with protection of anonymity, feasibility constraints, and the potential benefits to the public that may arise from a better understanding, or detection of mental health status (a perspective that may be informed by public health ethics [35]).

Nevertheless, there may be a need for more explicit dialogue and efforts to nurture a clearer understanding for those from whom data is being collected and analyzed as to what constitutes the purpose of the data analysis, and what risks and benefits sharing their data may entail for the

Table 3. Types and Frequency of Ethical Issues or Approaches that Were Described or Addressed in the Papers

	Detail/Steps Taken	Paper/Author(s)
No mention of ethics/ethical concerns	N/A	Adamou et al. [2]; Arguilar-Ruiz et al. [4]; Alam et al. [6]; Chang et al. [31]; Chen et al. [33]; DeMasi and Recht [44]; Diederich et al. [45]; Fatima et al. [57]; Frogner et al. [62]; Galiatsatos et al. [63]; Gjoreski et al. [67]; Joshi et al. [86]; Kavuluru et al. [89]; Mallol-Ragolta et al. [112]; Nguyen et al. [133]; Nosakhare and Picard [135]; Panagiotakopoulos et al. [139]; Patterson and Cloud [144]; Pestian et al. [145]; Rastogi et al. [154]; Ray et al. [155]; Spathis et al. [177]; Tavabi [184]; Tran et al. [192]; Tsiakas et al. [193]; Wilbourne et al. [205]
Reports of ethical approval/review exemption	Institutional/Regional IRB	Doryab et al. [50]; Ernala et al. [53]; Nobles et al. [134]; Paredes et al. [140]; Salekin et al. [168]; Wang et al. [201]; Yazdevar et al. [211]; Zakaria et al. [217]
	Re-use of data (e.g., that previously received or was exempt from ethical approval)	Feng et al. [61], Gaur et al. [64]; Morshed et al. [128]; Quisel et al. [152]; Spathis et al. [176]
	Statement of having "ethical clearance"	Ojeme and Mbogho [136], Srividya et al. [179]
	Statement of study and data being exempt from ethics review	Park et al. [141]
Privacy protection	Public data access + user anonymization	Park et al. [141]; Saha and De Choudhury [165]; Yazdevar et al. [211]
	No recording of person identifiable data (e.g., text, speech, low-level interactions)	Cao et al. [27]; Mitra et al.; Morshed et al. [128]; Rabbi et al. [153]; Salekin et al. [168]; Zakaria et al. [218]
	Confidential treatment/no (public) sharing of data	Salekin et al. [168]; Wang et al. [201]; Zakaria et al. [218]; Zhou et al. [222]
Consent and user control over data use; ability to contest ML	Informed consent prior to study for primary data collection	Broek et al. [23]; Feng et al. [61]; Srividya et al. [179]; Yang and Bath [209]; Zhou et al. [222]
	Need for users to choose data source used for diagnostic assessments	Jain and Argawal [82]
	Ability to contest system feedback	Hirsch et al. [77, 78]; Nobles et al. [134]
Study planning and conduct	Study risk assessment/planning with, or supervision by MHP (e.g., licensed clinical psychologist, practicing psychiatrists)	Nobles et al. [134]; Salekin et al. [168]; Zakaria et al. [218]
	Study coordination by person trained with relevant expertise	Broek et al. [23]; Nobles et al. [134]; Salekin et al. [168]
	Post-study mood assessment to identify/help mitigate any induced negative experiences	Nobles et al. [134]
	Avoidance of mental illness screening or specific data instruments to avoid harm	Paredes et al. [140]; Zakaria et al. [218]
Broader implications and guidelines	Broader impact of interventions (e.g., on health work-practices, patient well-being)	Ernala et al. [53]; Hirsch et al. [77, 78]; Zakaria et al. [218]
	Justice/fairness	Zakaria et al. [218]
	Lack of ethical guidelines/data regulations	Chancellor [28]; Morshed et al. [128]; Zakaria et al. [218]

person. This could be crucial for supporting people's autonomy and their ability to make well-informed choices.

Finally, the concept of autonomy also needs to be considered where ML model results are used as part of interventions that could drive or automate (clinical) assessments and decision-making processes. In Section 3.6.3, we described findings by Hirsch et al. [78] that showed a tendency by MHPs to perceive ML system evaluations as "more objective," and to be over-trusting of ML outputs—irrespective of a clear understanding of how results were derived, nor if they were accurate. This over-reliance however can have strong negative implications if model predictions are wrong, and difficult for people to scrutinize or contest. Nobles et al. [134] exemplified this through perhaps an extreme example that raises awareness how—in the context of a false ML alert of high suicide risk—peoples' autonomy could be claimed by health care services. Reflecting on questions of care responsibility, and how system outputs may become evaluated by MHPs, and compared with human judgement if the person denies the result, the authors [ibid] write: "*The field would need to answer questions related to mandated reporting and involuntary hospitalization. For example, would a clinician be legally and ethically mandated to intervene as they would if a patient endorsed active suicide intent in person? What is the most appropriate action for someone who denies having suicidal thoughts, plans, or intent but whose text messages indicate elevated risk?"* (p. 7).

Again, this emphasizes the need to better assist laypeople in evaluating the capabilities and limitations of ML models to help counteract tendencies to uncritically accept machine-generated insights (cf. [78]).

3.6.4.2 Beneficence and Non-maleficence. All papers were motivated in their work to positively contribute to mental health and peoples' welfare. The principle of beneficence however does not only entail encouraging human flourishing and well-being by doing the right thing, but also suggests to "do it well" [13]. This means that ML-enabled mental health interventions should be designed to maximize benefits and minimize harm (cf. [170]).

Most explicit considerations of non-maleficence were apparent in a few works (n = 5) that described active approaches to avoid "harm for study participants" as part of data collection efforts. This includes the *joint planning and assessing of risks involved in data collection studies together with MHPs* [134, 168, 218] and the *presence of a trained psychologist or therapist* during research activities to safeguard participants who may experience distress [23, 134, 168]. In addition, some researchers made explicit choices to *not screen for the presence of any mental illness* [140], or to *omit critical clinical questions such as "item 9" on the PHQ scale* that assesses suicidal thoughts [218]. Regarding the latter, the researchers acknowledged that a non-clinical research team may lack the necessary training to handle any definite answers to this question [ibid].

Outside of user study reports, there was a lack of critical engagement with the potential implications of introducing generated ML outputs into real-world mental health or care practices. Although the papers described excitement with how achieved ML models and related insights might come to benefit people, there was little reflection on *how* people might respond to systems that identify, or "diagnose" them with a mental health problem; and alert them, or others, of specific "risks." For example, Zakaria et al. [218] describe the possibilities of applying their ML system as an intervention as follows: *"In real-world operation, students who are concurrently depressed and severely stressed and frequently depressed but not severely stressed are those that StressMon detects as 'red-flags' so that interventions can take place as early as possible"* (p. 13). The authors however also report a misclassification rate of 18.20% in stress detection, which meant that most of their participants were identified by the system as "severely stressed" at some point in the study; and for 9 (out of 55) students there were several instances of depression misclassification, including one student who remained completely undetected by their model [ibid]. What this and similar works therefore fail to acknowledge or discuss is how proposed classifiers may come to be sensibly implemented in practice, and what the risks and implications might be of such interventions; especially when model predictions are likely false at least some of the times.

In another instance, Chen et al. [33] described the implications of their work on predicting Twitter users with symptoms of depression from self-report diagnosis posts in this way: *"After learning the traces and patterns of depressed users from these features, the trained classifiers can be easily applied for detecting Twitter users with depression who did not post about their conditions and users who are at risk of depression"* (p. 1660). Related to this instance, open questions remain about the extent to which individuals may appreciate or reject the idea of "depression detection" from their Twitter uses, and what harm could arise if communications of such a proxy diagnosis are not carefully scaffolded and appropriate safeguards in place to support the person. There is also concern about potential uses of such technologies to deliberately identify and target individuals who may be more vulnerable (e.g., with advertising).

Thus, it is important for researchers to have awareness and recognize potential risks of harm that may come from how developed ML insights or systems may be applied and appropriated in practice. In our corpus, Hirsch et al. [77, 78] are among the very few that considered the broader use and potential negative implications of ML predictions within a specific health care context. They described, e.g., risks of supervisors of mental health counselors, who are being assessed and "judged by a machine," to potentially rely too heavily on ML recommendation in evaluations of job performance. The authors also warned about risks of increasing financial and organizational pressures to "rationalize" mental health care through ML technology; as well as counsellor concerns about workplace surveillance and decisions to "fire" someone based on automated skill assessments. As a result, trainee therapists could potentially start adapting their practices to improve machine scores rather than their counselling skills, which, ultimately, could be detrimental rather than helpful to patient care. These examples foreground the importance of a more critical engagement with the broader ethical, societal and workplace challenges that can be bound up with new ML systems.

3.6.4.3 Justice. Finally, the ethical principal of justice focuses on the fair distribution of benefits, risks, and costs [170] and is often treated synonymously with fairness [39]. In the context of ML research, this can include the study of what constitutes a fair distribution of resources in the design and evaluation of algorithmic systems; removal of bias from the ML learning process (see [53, 78, 141], Section 3.5.2); or the perceived fairness of a decision-making process [102]. Only one paper [218] explicitly mentions "justice," and describes it as requiring fair participation: "*Fairness is true for StressMon, as its data collection is not influenced by factors such as the socioeconomic status or technical experience of the user. Instead, StressMon leverages Wi-Fi, which is readily available in public spaces (*e.g.*, offices, campuses and shopping malls) and commodity devices (*e.g.*, laptops and mobile phones)"* (p. 23). Here, it is argued that fairness is ensured since the resource provided—an infrastructure system to monitor stress and depression—is available to all people through their devices. What's missing in such arguments, is the acknowledgement that not all people may have access to, or continuously carry, laptops or mobile phones (e.g., due to the digital divide [70]).

4 DISCUSSION

This systematic review provides an introduction to the emerging area of research and development of ML in mental health. We now discuss existing approaches and future directions based on three key trends and associated challenges that we identified through this review: (i) identifying important health care needs to inform ML development, (ii) evaluating the effectiveness of

ML-interventions, and (iii) understanding the broader implications of new ML systems through deeper study within real-world contexts.

4.1 Identifying Key Health Care Needs and Problem Definitions for ML

The findings of our review show a recent growth in ML research in the domain of mental health, with many of the works seeking to explore how ML could be leveraged for "social good" by helping to address the significant personal and economic burden that is caused by mental illness. In line with recent reports by Shatte et al.'s [173], the vast majority of this research described approaches to the detection and diagnosis of mental health behaviors or conditions. Fewer works explored how ML approaches can support our understanding of mental health (e.g., [28, 89, 133, 141, 165]) or be leveraged in treatment (cf. [4, 77, 78, 140]). This raises the question how meaningful research questions and problem scenarios for ML are commonly identified, and how best to support such choices to maximize ML utility in the mental health domain.

Here, one assumption might be that the general need for access to large-scale, high-quality mental health data required for ML modeling plays a moderating role in the types of research questions and ML applications that are being developed. In the health care domain in general, and for mental health specifically (e.g., [134]), there is an emphasis on the challenges and costs that are involved in gaining access to, and collecting data both at scale, and of sufficient diversity. As shown in our analysis (Table 2), and especially for data collection studies, the numbers of participants that represented patients or people with a mental health condition was often low —especially when considering the data demands of advanced ML techniques. Larger numbers were achieved in analysis of health records, yet their access is often restricted to, and requires collaboration with, health organizations. As a result, there is a risk that the expense of data collection may limit study design, forcing researchers to use readily available data (e.g., social media, public databases). Such data, in turn, may be suboptimal for exploring a particular research question outside of the original data context. Similarly, the availability of clinical outcome measures to assess mental health through clinically validated scales and screening questionnaires (Section 3.3.2) may also contribute to explanations of the prevalence of algorithmic modeling to assist particularly in mental health symptom detection and diagnosis.

We believe, however, there is a lot more scope for other, perhaps more important and innovative uses of ML if we were to ask: how ML can meaningfully augment existing health care practices, or help make certain processes easier or more effective for mental health service users. *Finding the most beneficial (as opposed to the most obvious) applications of ML* will require creative exploration of the design space coupled with an *understanding of the real problems faced by potential users and mental health services* on a day-to-day basis. Next, we (i) expand on proposals to identify key mental health care needs and broaden the focus of ML and (ii) suggest more active, yet careful approaches in negotiating data access to lift constraints.

4.1.1 Wider Opportunities for ML: Moving Beyond Mental Health Detection and Categorial Diagnosis. A key motivation of the majority of review papers was the development of ML models to help achieve more effective tools or approaches to aid mental health assessment and monitoring. As a new and evolving area of research, there are however a lot more opportunities for ML to expand the scope of what is currently possible.

Understanding Mental Health Status and Discriminating between Disease Categories. Thus far, few studies have sought to advance our *understanding of mental health* by extracting the importance of identified (behavioral) features, their combinations, and relations with mental health [50, 57, 63, 89, 133, 135, 139, 141, 165, 201, 209]. The vast majority of papers described ML classification tasks aimed at identifying whether a particular individual belongs to a particular diagnostic category, e.g., "depressed" or "not depressed." However, looking at a mental illness, like depression, as one broad category may not take the variability of depression symptoms into account, and how the illness manifests [189]. Furthermore, in the medical domain, and for everyday psychiatric practice, it is often argued that the more challenging question is often not detecting the presence of mental health conditions and whether a person is in need of treatment, but the *differential diagnoses that discriminate between multiple likely illness categories, and to identify optimal treatments* [25]. Here, ML approaches such as multi-class prediction or multi-task learning may be well suited to explore differences across mental illness subtypes or treatment groups. ML techniques may also assist in *identifying yet-to-be-discovered mental illness dimensions* and support recent clinical efforts that seek to supplement discrete definitions through a *more continuous, dimensional symptom system* [ibid].

Personalizing and Optimizing Mental Health Treatment. In our review corpus, one paper explored how ML could enable *personalized recommendations* for stress treatment [140]. There is ample scope for future work to study how ML could be applied to allow, e.g., for a *more effective tailoring of interventions to each persons' unique mental health and support needs*; and *assist in the development of more effective mental health treatments*. The ability to potentially predict treatment effectiveness on an individual level presents a particular benefit of ML approaches over traditional clinical and statistical methods, whose aim often is to identify treatment options that explain the benefits and variance for the "majority of a clinical group," and formally test for "group effects" (cf. [25]).

ML approaches also have potential for enabling more targeted adjustments to treatment through advancing our understanding of what types of interventions, or their form of delivery, may work most effectively for particular people [11, 172, 194]. Albeit still scarce, studies are starting to emerge that propose uses of ML to provide just-in-time adaptive interventions (JITAIs) (e.g., [84, 130]). Often motivated to create more engaging, responsive and adaptive treatments based on information about the person or their environment, JITAIs utilize algorithms to optimize interventions for each person based on proximal outcomes [172]. For example, Jeong and Breazeal [84] employed ML to assess a persons' emotional state (analyzing their facial expressions and SMS) and used this information to tailor what positive psychology intervention the person would receive.

For digital mental health services specifically, such as online cognitive-behavioral therapy (iCBT) interventions (e.g., SilverCloud Health[10] [49, 127, 157]) or mobile mental health apps (e.g., IntelliCare[11] [125]), there is further great potential for the analysis of log event data [128, 194]. For example, ML could be used to discover usage patterns in log data that can help predict future user behaviors or mental health states [194]. This may include predictions of *users' risk of drop-out from treatment*, or *risk of rapid declines in mental health* through which more timely and bespoke interventions could be enabled. Other approaches, such as association analysis, can further help uncover what features in a digital behavioral health intervention often occur together [ibid] and help derive opportunities for personalization and to optimize treatment. This has recently been exemplified by Chikersal et al. [34], who used association rule mining (ARM) to learn what about the communications of therapeutic supporters who guide patients through an iCBT program for depression and anxiety is linked with better improvements in patient mental health. Specifically, the authors analyzed how specific linguistic strategies in support messages to patients correlated with better patient outcomes *dependent on the patients' specific context* (e.g., their current mental health, treatment week, level of engagement with iCBT). The research showed how certain support strategies (e.g., use of more positive words, or words referencing social behaviors) were "more" or

[10]https://www.silvercloudhealth.com/.
[11]https://intellicare.cbits.northwestern.edu/.

"less" important depending on how actively users engaged with the treatment. This, in turn, can help human supporters of *iCBT* interventions to better tailor their communications to each clients' circumstances. Explorations of ML use for assisting the communication skills and work practices of MHPs have also been evident in a small number of papers in our review corpus in the context of *face-to-face therapy* [77, 78] and for improving coaching via a *text-chat app* [205].

Supporting Positive and Preventative Approaches to Mental Health. Lastly, we want to note that the vast majority of paper records focused on symptoms and conditions indictive of mental health difficulties. This leaves scope for uses of ML in *supporting preventative approaches* (outside of acute risk detection) and *assisting in positive mental health outcomes* (e.g., resilience, self-determination, personal growth). Under-exploration in these areas may partly be reflective of a lesser understanding of "positive mental health" concepts [188], and a lack of available data [173]. This underrepresentation may also be partly due to the search methodology applied in this review, which did not include terms like "mental well-being," "psychological well-being," "subjective well-being" or related constructs.

4.1.2 Data Access Challenges: Identify Tradeoffs for Data Sharing that People are Willing to Make. Bound up with challenges in identifying important health and care needs are requirements for access to relevant, large-scale, high-quality data to allow for effective ML modeling. This can be particularly difficult in the mental health domain due to ethical and privacy challenges involved in (i) recruiting individuals who may be more vulnerable to research [132] and (ii) the time-consuming and effortful nature of data acquisition that often requires multi-disciplinary partnerships with health care providers and do not scale easily [53].

Improving Informed Consent Processes and Users Trust in Data Applications. For many of the social media studies papers, the pooling of "publicly" available data [e.g., 141, 165, 211] has often been described as not requiring explicit user consent. Recently, there is however increasing debate on whether the use of public data to predict, e.g., mental health states, may border on medical diagnosis and should be considered as human subjects research [29]. This is echoed in user research that suggests that social media users often do not have awareness that their online content is used for research, and express concerns about such use "without their consent" [58]. Describing how people attitudes to data use are highly contextual, Fiesler and Proferes [58] found that Twitter users "felt less comfortable" about uses of their entire Twitter history (rather than individual tweets), and where content had more personal significance or sensitivity. They also described ideas of data consent or permission as stemming from the underlying importance of *respect for the user* and the need for data uses (for research or ML applications) to *align with user expectations*. Although obtaining consent at scale presents a practical challenge [29, 165], there are increasingly proposals for how users could, at least, be informed about the use of their data and be given opportunities to opt-in or opt-out (e.g., by tweeting that their tweet is included in research) [58]. The feasibility of such approaches will require future testing.

This example, and the need for access to rich, personal data for developing effective ML models and interventions, also raise the question how to ensure that people generally agree to, and can trust researchers and data applications with the collection and processing of their sensitive information? Likely, *this requires careful tradeoffs between data needs for algorithmic purposes and how related data practices are justifiable in terms of benefit or potentially harm to the person* (cf. [119]). For example, while sensitive data such as a person's gender, age, or clinical diagnosis can aid in differentiating health-behavior patterns and groups, and enable testing for diversity in a dataset [66, 76, 80]; we have to consider how comfortable people may feel about providing such data. For this, individuals need to be better supported in assessing "the potential benefits of data sharing" and "how potential risks are mitigated through safeguards or outweighed by the potential benefits" (e.g., effectiveness of interventions). This will enable them to make more informed choices about data uses; and, in turn, aid their trust in, and acceptance of, data applications.

This might be achieved by (i) making processes of how we seek consent more comprehensive and usable (in line with GDPR regulations), (ii) explaining more clearly the benefits of data use to the person and the mechanisms employed to protect their data (taking active steps to mitigate risks), and (iii) ensuring that people have more control and actual choice(s) about whether their data is being used for specific ML purposes or not.

Need to Develop Responsible Approaches for Data Sharing and Data Donation. Difficulties in gaining access to mental health data have also led to proposals to build and leverage shared infrastructures and data repositories for conducting data research [53]. Creating benchmark datasets [73] and having better methods for data sharing can support the replication of research findings and improve scientific quality [124]. For example, systems such as the Clinical Record Interactive Search (CRIS) enable researchers to access large-scale electronic mental health record data from the UK. To ensure responsible use, applications for data access are reviewed and monitored by a committee for compliance with ethical and legal requirements [116]. Similar initiatives exist in the US, e.g., through the Connected and Open Research Ethics (CORE) program that manages shared health care resources and helps navigate many of the complex, ethical, and practical challenges involved in collecting sensitive health care data [191]. Other data collection efforts include crowdsourcing and data donation programs such as PatientsLikeMe[12] and OurDataHelps,[13] where people can choose to share data and information about their health for data science and research purposes. Research charities are also playing an emerging role by matching researchers and their research questions to datasets [116] and providing funding for mental health research (e.g., MQ charity[14]).

4.2 Evidencing the (Real-World) Effectiveness of ML-Interventions

With few exceptions [77, 78, 140], the papers in our corpus primarily assessed the effectiveness of newly developed ML models based on their predictive performance—measured in terms of accuracy and errors (cf. [149]) and comparison with (state-of-the-art) baselines—on held-out data. Yet, this often provides little insight as to how reliably a model may perform in the real world; or how it would find useful adoption within health care services. As such, these papers predominantly provide proof-of-concept studies that necessitate continued research and development to further improve (classification) accuracy [173]. Further, there is little exploration of how developed ML approaches would be perceived by, and come to actually benefit, their proposed users (e.g., clinicians, patients, online community moderators).

4.2.1 Beyond Accuracy in Model Performance: Risks of Overclaiming and Premature Generalization. As is perhaps less surprising in a review of the computing and HCI literature on "ML applications" in mental health, we found that the majority of papers focused on the *technical or algorithmic development of initial ML models* (Section 3.5.1). As such, they predominantly report their technical contributions through new data methods and accuracy metrics (Section 3.4.2) and discuss key data modeling challenges (Section 3.5.2). At the same time, many of these technical papers also include speculative descriptions, proposals, and claims as to how their new ML models may come to be used to impact clinical or everyday health management practices (Section 3.6.1). Despite great enthusiasm for how ML approaches could be transformative to the mental health domain, *it is important to not to prematurely overclaim anticipated (clinical) benefits or generalize*

[12] https://www.patientslikeme.com/.
[13] https://ourdatahelps.org/.
[14] https://www.mqmentalhealth.org/.

too soon from initial proof-of-concepts. Next, we discuss the importance to (i) acknowledge how the conduct and impact of research and technological development is assessed and shaped by different scientific disciplines and (ii) be cautious in making clinical or diagnostic claims where datasets did not include much, or any, "clinically validated" data (cf. [64, 128]), and where achieved ML model results were not evaluated or studied in actual health care contexts.

Acknowledging and Addressing Disciplinary Differences in the Conduct and Evaluation of Research. In computing, research is typically exploratory in nature and seeks to "find" an answer to a question or problem. In contrast, clinical research tends to be hypothesis-driven and involve the design of studies to test and "confirm" an answer to a question [124]. Furthermore, while computer scientists often focus on proof-of-concepts (e.g., "Does it work at all?"), clinical scientists value generalizability (e.g., "Does it work for all populations at all circumstances?"). Often in a quest to identify "novel solutions," computing scientists can also have a higher tolerance for risks than clinical researchers, who value internal validity and confidence in research results [ibid]. Naturally, these disciplinary differences are reflected in the types of data sources that are used for ML analysis (e.g., clinical vs. general population/proxy data) as well as the methods that are employed to evaluate the "success" of the research or development output. This variability complicates the comparison of findings across studies [124]. It also means that for ML research that seeks to inform clinical diagnosis and decision-making, it is imperative that algorithmic models are built on (clinically) valid data [53], perform robustly and reliably outside their training or test environment (and without discriminating against sub-groups), and assessed for their practical usefulness and the value they might bring to real-world health care practices (e.g., reduced clinician time, improved patient outcomes).

Mental Health Constructs and Clinical Validity of ML Results. A significant proportion of the data collection studies (21 out of 29) did not include any patient data or external assessments by clinicians and were often conducted as part of lab or pilot studies that used frequent EMA's to gather "ground truth" data (e.g., a person's mood captured by an Affect Grid [176]). Data collected in this way can differ significantly from standardized clinical screening or assessment methods that are administered by trained MHPs. For example, in social media analysis, there has been an increase in criticism [29, 53, 165] of the use of self-disclosed, sentence-based labelling such as *"I was diagnosed with..."* [33, 105, 211, 219] as a mechanism for "diagnostic" ground truth, as this does not conform with clinical assessment tools such as the DSM [8]. The DSM provides a written manual for making accurate psychiatric diagnosis that is based on 60 years of empirical results [29]. Concerns about a lack of clinical grounding, theoretical contextualization, and psychometric validity were particularly prominent in the paper by Ernala et al. [53]. Their study compared different approaches to diagnosing social media users with "schizophrenia" and found poor external validity where ML models that were based on "proxy information" were tested on people who had a clinical diagnosis. Additionally, Chancellor et al. [30] also raised concern that many "mental health status observations" tend to be based on single units of observation (e.g., an online post) without additional context about an individual or any methodological substantiation of how a single moment of distress may relate to the presence of a mental health condition. Many social media studies further imply experimental rigor by including "control" groups. However, these are often selected as a random sample of online service users (e.g., [33, 211]), without any (formal) validation that these were individuals who did not have specific mental health symptoms (e.g., [30]). Outside of social media data, Saeb et al. [164] also called for caution in the interpretation of ML outputs following their review of studies that used smartphones and wearable sensors to predict clinical outcomes based on a publicly available dataset. Having replicated the approaches taken, they found that almost half of the examined studies used a popular cross-validation method (record-wise cross-validation) that significantly overestimates the algorithms prediction accuracy.

Thus, for developing effective and implementable ML systems for mental health, and as ML models advance in technical development and accuracy, more research is needed to (i) test the validity of the mental health constructs that are assessed (e.g., diagnostic validity) and (ii) ensure that ML outputs are transferable and their prediction robust for use in "practice" (reliability). Furthermore, as ML model insights are intended for use and become incorporated in real-world mental health interventions, future studies have to start assessing (iii) their practical use, value for, and acceptance by, key stakeholders (cf. [124]); as well as (iv) their actual effectiveness for improving (promised) mental health outcomes, and reducing costs. In this regard, it is recommended to involve MHPs and the individuals targeted by ML predictions throughout the research design and development process. Clincial experts, i.e., can provide key insights into construct validity, assessments of ground truth and biases, as well as important context information that can help in the interpretation of data findings, improve rigor, and manage deployment risks and tradeoffs [29, 53].

4.2.2 Avoid Dehumanization and Undermining the Value of Other Data Methods or Clinical Expertise. To evaluate the potential usefulness of new ML approaches, it is important to examine how existing work positions itself and its contributions to mental health research and practice. Section 3.2 described key motivations for the use of ML for mental health to include (i) unobtrusive or continuous data access, (ii) automatic data processing for efficiency and cost savings, and (iii) claims that data-derived assessments provide objective, more accurate and reliable information to help improve existing (clinical) tools and decision-making practices (e.g., [61, 192]). The latter argument in particular was often substantiated through an emphasis on the *disadvantages and insufficiencies* of traditional questionnaires and self-report tools (e.g., [153, 211]) as well as clinical approaches (e.g., [2, 63, 82, 134, 136, 139, 144, 145, 179]). Together, these arguments suggest a potential superiority of, and possibilities for, new data tools to "outperform" existing approaches [192]. Next, we discuss reasons for why it may be advisable to exercise caution in positioning ML contributions in this way.

Overcoming Methodological Limitations and Improving Insights by Combining the Strengths of Different Data Methods. Different research and data analysis methods contribute different types of insights and have their own strengths and limitations. For example, validated clinical tools present instruments that have been extensively tested psychometrically to ensure results are both accurate and consistent. The accuracy and reliability of ML models is, inevitably, limited by the quality of the data used in their training. ML models are also prone to *error, uncertainty, and bias* (cf. [190]). Even where ML models perform with minimal error, challenges remain for their generalization to contexts outside the specific training environment [106]. Taking these and other described data limitations (see Section 3.5.2) into consideration, and outside of much evidence of real-world evaluations of the effectiveness of enabled ML predictions, researchers working in this space *need to be careful with any claims that data-derived assessments indeed provide more "objective, accurate, and reliable" information.* We believe, in these early stages of research and development, that it is important to set and communicate appropriate expectations of what ML outputs, to-date, can realistically achieve and what their limitations are. This is particularly important for setting up successful research collaborations and productive ML development partnerships with health care providers. Here, a more open dialogue about the potential and challenges of achieving robust ML models is important for nurturing empathy and trust. This can pave the way for health care providers to better comprehend what is required of them, for example, to ensure "good-quality data capture" as well as developing their understanding of the unique data analytics capabilities that new ML approaches afford. For example, a key strength of ML methods is their capability to mathematically identify, e.g., the most relevant variables in a dataset based on an outcome of interest. In contrast, conventional statistical methods typically rely on the investigators—their assumptions and expertise—to specify the variables that are relevant for a particular analysis [100, 194]. Similarly, while studies such as randomized controlled trials (RCTs) can have advantages in helping to control for, and reduce certain sources of bias when assessing the effectiveness of an intervention, they reveal little insight as to why or how certain factors contributed (more or less) to an outcome (cf. [194]). All this suggests the *need to better understand what different research and data methods can explain and contribute to knowledge generation*, and *how they could best come to complement (rather than compete with) each other.*

Empowering MHPs through Data Insights and Supporting Their Agency as Health Care Experts. Much of the reported ML work is motivated to help develop new tools and methods to assist mental health care, which is often provided through MHPs. Therefore, it is important to be careful in the positioning of new data methods or systems to not unnecessarily undermine the important role of health or care providers. This can risk reducing their willingness to support the development as well as adoption and acceptance of ML approaches into their work practices. Instead, future work should explore *how to design ML-interventions such that they can become valuable tools to assist clinicians in their information needs and decision-making processes rather than attempting to replace or outperform them in their clinical expertise.*

Avoiding Stigmatization and Dehumanization. In their recent review of ML approaches used for mental health predictions in social media, Chancellor et al. [30] critiqued how humans were represented in data research. In various studies, the authors found that individuals who may not have a mental health condition were often described as "normal" or "neurotypical." Such terminology however risks stigmatizing people who have a mental health condition by *othering* them and their experiences. Further, the authors described trends in computationally focused work to treat individuals as "data points" for machine training and optimization. At an extreme, humans become the "objects" of analysis and represented through their online "accounts" and "blogs." In dividing the person from the data, unique details of their experiences are abstracted away to identify large-scale patterns or phenomena. Such simplifications are at odds with the complexities and subtleties of people's lives and felt mental health experiences. While HCI research tends to place the human and their needs—as the "subject" of interest—at the center of technology design processes, the area of ML—drawing on statistics, computer science and optimization research—views the abstracted model or data point as the "object" of study. Yet, within human-centered research, such objectifications can imply a stronger interest in machine analysis than the people that the research suggests it is interested to help. As a consequence, this can potentially cloud the responsibilities and ethical priorities of the researchers [30]. Thus, it is important that researchers are mindful in their reporting practices to *avoid stigmatizing language that can harm people and* communities; and *diminish objectification* as people and their individual experiences are being transformed into compressed mathematical representations.

4.3 Understanding Opportunities and Risks of ML-Systems in Context

As evident in this review, the field of ML in mental health presents an emerging area that, so far, has mostly contributed to the discovery and development of basic (multi-disciplinary) research insights, with very few initial investigations of potential ML interventions (cf. [77, 78, 140]). The field of Implementation Science often describes this as the initial stage in what presents a complex, multifaceted process of moving important research innovation into actual work flows, and for sustaining and scaling-up effective health care interventions [68]. To endure on this early journey toward achieving real-world impact, researchers in HCI and ML will need to (i) continue basic research, (ii) expand development and initial testing of new ML-interventions, and (iii) start moving towards a rigorous analysis of the effectiveness of these interventions. To further maximize the usefulness of potential ML interventions, it is paramount to (iv) more actively include MHPs and people with lived experiences of mental health in research and development processes (e.g., through observational studies, interviews, focus groups, and collaborative partnerships), and to develop and study new ML techniques in real-world settings. Although the majority of review papers presented contributions by multi-disciplinary research and engineering teams, there was little reporting of user-centered research methods and explicit dialogue with MHPs and other potential beneficiaries (outside of data access and labelling efforts) to inform ML research. A stronger collaboration with, and closer involvement of, key stakeholders will be crucial to identify important health care needs and scenarios for ML development. Simultaneously, the study of new ML systems will likely foreground new challenges (e.g., adoption into work practices, ethical issues) that will need to be considered and addressed if ML-enabled interventions are to succeed in the real world (cf. [20, 79, 173]).

There remain many challenges in order to move from proof-of-concept explorations toward the design and study of ML tools and interventions that are useful in broader populations [124]. To help move toward this goal, our review foregrounded the following two areas of research and development that require further consideration: (i) the need to better support laypeople's understanding of ML outputs and (ii) to recognize and appropriately respond to broader practical and ethical implications that can be bound up with the use of these interventions in real-world mental health care contexts.

4.3.1 Design to Support Appropriate *Understanding and Use of ML-Outputs by Laypeople.* A key challenge in the design of ML-enabled systems is the generation of outputs that are interpretable and (clinically) useful to mental health care providers or target recipients. To address these challenges, work in this area often includes methods which support "understanding of the model" (e.g., [106, 149]). They include the following: extracting (and visualizing) model outputs and properties; estimating the influence of training examples; or learning local approximations to explain individual predictions of complex models post-hoc. Beyond this more data-driven understanding of methods, interpretability is mostly understood in terms of end-users being able to simulate, trust, or debug model decisions [1, 78, 149], and designing interactions with intelligent systems that can aid human understanding and decision making (cf. [20]). In our corpus only one paper [78] engaged explicitly with this topic of *ML intelligibility*. The authors [ibid] described the tendency of participants to perceive and uncritically accept ML-generated outputs as "factual information," even when they acknowledged that they did not fully understand how feedback was derived, or what it precisely meant. Their willingness to *trust* the ML system was found to be bound up with the perceived "legibility" of the output—the extent to which the system seemed to "make sense" to the user—rather than its statistical accuracy. This demonstrates the *need for more research investigating how an appropriate interpretation of ML outputs by laypeople can be supported* (see also recent work by [212]).

To support laypeople's understanding of how specific (behavior) data and model results relate to a health outcome, technical or mathematical explanations of model accuracy or uncertainty however might be limited. For example, how should MHPs interpret the significance of a prediction that indicates e.g., an 83% risk of suicide; and how can they meaningfully differentiate this from a 78% or 88% risk prediction? To enable and support laypeople to appropriately make use of ML outputs, user interface design and interactive visualizations or simulations can play a key role in generating comprehensive mappings for users, and help them assess, inspect, and cross-validate ML outputs in line with their own assessments of a situation. This is needed to better enable laypeople to calibrate their understanding of a system's capabilities and limitations to reduce risks of over-reliance on potentially over-confident predictions [109, 111]. To support scrutiny and encourage more careful interpretations of ML interferences, this suggests the need for (i) stronger efforts in supporting peoples' awareness of the probabilistic (rather than deterministic) nature of many ML models, and their likely proneness to errors, (ii) the provision of relevant additional context information and evidence that can help users to affirm, or contest ML outputs [77], and (iii) the inclusion of opportunities for user input and a strengthening of their role as data controllers [7] through encouragements to ask questions, to inspect any conclusions that seem unreasonable; and to facilitate the recording of any disagreements with a system (cf. [77]), or even correct identified errors. For early examples of HCI approaches to assist the interpretation and use of ML-enabled interventions, see recent work in health care more broadly, such as personalized fitness apps [52] or clinical support tools in critical surgery decisions [210].

4.3.2 Recognize and Respond to Broader Practical and Ethical Implications of ML-System Use. ML systems are increasingly becoming "real" [79] and embedded in high-stakes domains like mental health care [77, 78], where they can have significant implications for people's lives. Thus, we need to give close consideration to the practical challenges and broader, often un-anticipated ethical risks that can be bound up with the design and deployment of ML-interventions for mental health, and pro-actively work to mitigate risks associated with their use in practice.

Across the review papers, we found generally little discussion of ethical issues outside reports of formal research approvals, user study considerations, and the adoption of risk-averse and privacy protecting data management practices. Few papers engaged with the broader implications of using developed ML models within health care practice; mostly when reporting errors in model predictions [53, 134, 218]. This included discussions of: how ML systems may implicate the relationship between different stakeholders (e.g., patients, clinicians, supervisors) [78]; how ML algorithms misclassified or did not at all detect certain individuals [218], and how a false-alarm of a high suicide risk alert, could—at an extreme—lead to a person's involuntary hospitalization [134]. Even in less extreme cases, the false identification of a mental health condition could have severe implications for a person's self-esteem, reputation or employment (particularly for people working, i.e., in the police force, as pilots) [29, 189]. This raises the question of who is responsible and accountable for errors and for making choices if and how individuals should be alerted to their own mental health status [29]. Within the broader literature on HCI in digital mental health, researchers have started to discuss the challenges involved in making people aware of machine-detected problems in ways that are sensible, and respond carefully to peoples' expectations, needs, or troubles (e.g., [170, 216]). Such challenges are particularly prevalent in contexts where behavioral analysis is done outside of explicit user awareness (e.g., mental health inferences drawn from a person's social media). For example, Young and Garett [216] outlined a first working protocol that suggests when, and which stakeholders should intervene (or not), in the case that people were found to express suicidal intentions on social media. This, and similar works [29], acknowledge the need for researchers to have a process in place for supporting people who are identified as "at risk," including who to contact, and what information to share to address (psychological) concerns.

Thus, to help realize the potential of ML to truly benefit people, and find acceptance by them, requires researchers to (i) engage in more open discourse about the opportunities as well as ethical difficulties bound up with the use of ML for specific mental health contexts and (ii) extend efforts to collaborate more closely with health care users and/or providers throughout ML system design and evaluation processes [29, 30].

4.4 Limitations

The corpus included in this review is by no means complete, and new work constantly emerging. We acknowledge that the implications of this work are limited by our search methodology that was restricted to broad terms ("mental health" and "ML") as well as our record selection criteria. As such, our work excludes important research and ML development for neurological and neurodevelopmental conditions and may under-represent other mental health related works that focus, for example, on preventative and more well-being-centric approaches. The article also has limitations due its focus on the computing and HCI literature through the ACM Guide. This was a deliberate decision to provide a report that focused on the current landscape of computing research where ML has become applied in the context of mental health. It enabled the inclusion of more in-depth descriptions of existing research and development, as well as rich discussions of current trends and important challenges with regards to data access, conceptualization and modeling of mental health behaviors, and broader ethical and real-world implementation and use considerations. To reduce risks of bias in data collection (i) identified paper records were independently screened by two researchers and disagreements resolved through full-text review and discussion and (ii) a data extraction sheet was used to systematically elicit key information from each paper. Care was also taken in reporting the findings to balance the accounts of different approaches and findings; with reports and interpretations continually reviewed and re-evaluated by all members of the research team.

5 CONCLUSION

Recent years have witnessed an increase in excitement and exploratory research on potential applications of ML for mental health. Our review has offered an overview of this area of research and highlighted current trends and challenges. Aiming to shape the future direction of work, we have discussed current approaches and potential steps toward achieving ML systems that are effective and implementable for mental health care.

Specifically, we have examined how constraints and requirements for access to large-scale, high-quality data can pose challenges to study design and urge researchers to extend efforts to gain more in-depth understanding of the specific needs or challenges that are faced by MHPS and people with lived mental health experiences. Deeper and more creative explorations of the design space can meaningfully inform future research questions and problem scenarios for ML to ensure the domain can truly benefit from novel data tools. This may extend beyond more obvious ML applications for mental health. Bound-up with data access is the need to better assist people in assessing potential benefits of data sharing and how potential risks are mitigated or outweighed by potential benefits (e.g., effectiveness of interventions), such that they can make more informed choices about data uses and to aid their trust in, and acceptance of, data applications.

Furthermore, since the field of ML in mental health is still in its infancy, we have urged for more caution in presentations of ML development to avoid premature claims on the potential usefulness and real-world impact of new models. This is especially important considering the complexity and difficulties involved in generating robust as well as technically and clinically reliable ML outputs. So far, the majority of models are rarely tested for use in clinical environments, leaving gaps in assessments of their practicality, acceptance, and effectiveness for improving mental health-related outcomes, or services.

In addition, while it was often argued in the literature that novel ML models have advantages over existing research and clinical methods, we suggested to look at these as complementary approaches to knowledge generation. Furthermore, we proposed that there is a lot more scope for future research to also extend explorations of how ML-interventions can become valuable tools to address the needs not only of mental health care recipients, but to support the practices of mental health care experts. In applying ML approaches to the capture and assessment of rich human needs and experiences, researchers should also be mindful to not translate and abstract away too much from the individual person and their unique context in data analysis, interpretation, and representation.

Finally, we argued that helping the field achieve its many ambitious visions for ML in mental health requires continued efforts in conducting basic, multi-disciplinary research in deep collaboration with health partners, developing and testing new ML-interventions, and studying their effectiveness within real-world use contexts. This includes a key focus on the challenges of designing new ML-enabled systems that are sufficiently interpretable and (clinically) useful to its target users or recipients. It also requires that research and development efforts recognize and carefully respond to the broader practical and ethical implications that the use of ML systems could have for people, health care, and society.

A APPENDIX

A.1 Examples of Common ML Models or Techniques in Each ML Algorithm Category

	Supervised	Unsupervised	Semi-supervised	Novel Methods
ML models/ techniques	Support Vector Machines (SVM) *k*-Nearest neighbors (*k*-NN) Naïve Bayes (NB) Regression analysis, e.g., Logistic Regression (LR), Lasso Supervised Latent Dirichlet Allocation Decision Trees (DT) Random Forests (RF) Supervised Hidden Markov Models (HMM) Supervised Neural networks (NN)	*k*-means clustering Hierarchical clustering Unsupervised Hidden Markov Models (HMM) Latent Dirichlet Allocation (LDA) Unsupervised Neural networks (NN) Association rule techniques	Semi-supervised ML Self-training Mixture models Co-training + multi-view learning Graph-based methods	Deep learning (DL) Active learning, i.e., Reinforcement Learning (RL) Custom-ML methods

A.2 Frequency and Types of Specific ML/Data-Challenges and Limitation Described

Category	Subcategory	Detail	Paper/Author(s)
Capturing accurate/ reliable data	Need for ground truth, robust labels, and validation	No clear definition + reliable measure of subjective non-discrete experiences	Gaur et al. [64]; Gjoreski et al. [67]; Nosakhare and Picard [135]; Rabbi et al. [153]
		Challenges in generating low-dimensional, meaningful data labels	Kavuluru et al. [89]; Ray et al. [155]; Salekin et al. [168]
		Lack of clinical validation/information to infer mental health	Chang et al. [31]; Ernala et al. [53]; Nguyen et al. [133]; Saha and De Choudhury [165]; Salekin et al. [168]
		Ecological validity: Transferability of data (differences of lab- vs. real-world data)	Broek et al. [23]; Diedrich et al. [45: Study 1]; Gjoreski et al. [67]; Zakaria et al. [218]

Category	Subcategory	Detail	Paper/Author(s)
	Noisy/ ambiguous signals	Ambiguous words/lexical variations	Chancellor [28]; Nobles et al. [134]; Saha and De Choudhury [165]; Yazdavar et al. [211]
		Ambiguity in signals (e.g., for audio: robust speaker detection; distinguish personal speaking style from symptoms)	Chang et al. [31]; Mallol-Ragolta et al. [112]; Rabbi et al. [153]; Salekin et al. [168]; Spathis et al. [176, 177]; Zhou et al. [222]
		Managing irrelevant, redundant information	Ojeme and Mbogho [136]
Dataset limitations	Restrictions due to data subjects/scale/study context	Too small or restricted study sample/need for larger (more diverse) datasets	Adamou et al. [2]; Diederich et al. [45]; Feng et al. [61], Kavuluru et al. [89]; Morshed et al. [128], Nobles et al. [134]; Ojeme and Mbogho [136]; Parades et al. [140]; Park et al. [141], Pestian et al. [145]; Quisel et al. [152]; Ray et al. [155]; Salekin et al. [168]; Spathis et al. [177], Yazdavar et al. [211]; Zhou et al. [222]
		Unknown confounding variables + limitations of study context	Fatima et al. [57]; Saha and De Choudhury [165]; Salekin et al. [168]
		Reference dataset not explicitly designed for mental health-related analysis	Alam et al. [6]
	Biased, missing, incomplete data	General acknowledgement of biases inherent to model design and dataset used for training	Ernala et al. [53]; Hirsch et al. [78]; Park et al. [141]
		Difficulties due to missing data values/sparse data	Alam et al. [6]; Spathis et al. [176, 177]
		Need for inclusion of other information (e.g., biological and genetic data, fMRI, video, facial expressions, social media data)	Diedrich et al. [45]; Pestian et al. [145]; Mallol-Ragolta et al. [112]; Morshed et al. [128]
Data processing	Continuous data	Identifying optimal time-segments/features for data analysis	Frogner et al. [62]; Mallol-Ragolta et al. [112]
	Data integration challenges	Modeling multi-modal data of different signals, durations, densities; data fusion challenges	Cao et al. [27]; Mitra et al. [122]; Morshed et al. [128], Panagiotakopoulos et al. [139]; Ray et al. [155]; Rastogi et al. [154]; Tran et al. [192]; Zhou et al. [222]
		Complex mapping of multi-label classification	Ojeme and Mbogho [134]
Limitations of ML modeling/ implications	Modeling approach chosen	Detection of presence, duration + frequency of symptoms (not severity)	Yazdavar et al. [211]
		Use of retrospective data for predicting future behavior	Patterson and Cloud [144]
		Focus on population rather than individual	Doryab et al. [50]; Gjoreski et al. [67]; Nguyen et al. [133]; Quisel et al. [152]
	Claims	Limited ability to make causal claims	Morshed et al. [128]; Saha and De Choudhury [164]; Wang et al. [201]
	Errors	Errors in classifications/predictions and fallibility of ML Systems	Doryab et al. [50]; Ernala et al. [53]; Hirsch et al. [78]; Nobles et al. [134], Zakaria et al. [218]
Other	Need for data security	Secure storage and handling of data/need for secure models	Alam et al. [6]; Rabbi et al. [153]; Jain and Agarwal [82]; Wang et al. [201]

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. ACM, Paper 582, 18 pages. DOI: https://doi.org/10.1145/3173574.3174156

[2] Marios Adamou, Grigoris Antoniou, Elissavet Greasidou, Vincenzo Lagani, Paulos Charonyktakis, and Ioannis Tsamardinos. 2018. Mining free-text medical notes for suicide risk assessment. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence (SETN'18)*. ACM, Paper 47, 8 pages. DOI: https://doi.org/10.1145/3200947.3201020

[3] Ehsan Adeli, Kim-Han Thung, Le An, Guorong Wu, Feng Shi, Tao Wang, and Dinggang Shen. 2018. Semi-supervised discriminative classification robust to sample-outliers and feature-noises. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 515–522. DOI: https://doi.org/10.1109/TPAMI.2018.2794470

[4] Jesus S. Aguilar-Ruiz, Raquel Costa, and Federico Divina. 2004. Knowledge discovery from doctor–patient relationship. In *Proceedings of the 2004 ACM symposium on Applied computing (SAC'04)*. ACM, 280–284. DOI: https://doi.org/10.1145/967900.967960

[5] Malik Anas Ahmad, Nadeem Ahmad Khan, and Waqas Majeed. 2014. Computer-assisted analysis system of electroencephalogram for diagnosing epilepsy. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR'14)*. IEEE Computer Society, Washington, DC, 3386–3391. DOI: https://doi.org/10.1109/ICPR.2014.583

[6] Md. Golam Rabiul Alam, Eung Jun Cho, Eui-Nam Huh, and Choong Seon Hong. 2014. Cloud based mental state monitoring system for suicide risk reconnaissance using wearable bio-sensors. In *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication (ICUIMC'14)*. ACM, Paper 56, 6 pages. DOI: https://doi.org/10.1145/2557977.2558020

[7] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120. DOI: https://doi.org/10.1609/aimag.v35i4.2513

[8] American Psychiatry Association. 2019. *Diagnostic and Statistical Manual of Mental Disorders (DSM–5)*. Last retrieved 7 th July 2019 from https://www.psychiatry.org/psychiatrists/practice/dsm.

[9] Elena M. Andresen, Judith A. Malmgren, William B. Carter, and Donald L. Patrick. 1994. Screening for depression in well older adults: Evaluation of a short form of the CES-D. *American Journal of Preventive Medicine* 10, 2 (1994), 77–84.

[10] Mohammad R. Arbabshirani, Sergey Plis, Jing Sui, and Vince D. Calhoun. 2017. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage* 145, PT B (2017), 137–165. DOI: https://doi.org/10.1016/j.neuroimage.2016.02.079

[11] Greg Barish, Hilary Aralis, Eric Elbogen, and Patricia Lester. 2019. A mobile app for patients and those who care about them: A case study for veterans with PTSD + Anger. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth'19)*. ACM, 1–10. DOI: https://doi.org/10.1145/3329189.3329248

[12] Jakob E. Bardram, Mads Frost, Károly Szántó, Maria Faurholt-Jepsen, Maj Vinberg, and Lars Vedel Kessing. 2013. Designing mobile health technology for bipolar disorder: A field trial of the monarca system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*. ACM, 2627–2636. DOI: https://doi.org/10.1145/2470654.2481364

[13] Reid Bates. 2004. A critical analysis of evaluation practice: The Kirkpatrick model and the principle of beneficence. *Evaluation and Program Planning* 27, 3 (2004), 341–347. DOI: https://doi.org/10.1016/j.evalprogplan.2004.04.011

[14] Matthew J. Bauman, Kate S. Boxer, Tzu-Yun Lin, Erika Salomon, Hareem Naveed, Lauren Haynes, Joe Walsh, Jen Helsby, Steve Yoder, Robert Sullivan, Chris Schneweis, and Rayid Ghani. 2018. Reducing incarceration through prioritized interventions. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS'18)*. ACM, Paper 6, 8 pages. DOI: https://doi.org/10.1145/3209811.3209869

[15] Victoria Bellotti, and Keith Edwards. 2001. Intelligibility and accountability: Human considerations in context-aware systems. *Human-Computer Interaction* 16, 2–4 (2001), 193–212. DOI: https://doi.org/10.1207/S15327051HCI16234_05

[16] Charles C. Benight, Kotaro Shoji, Lori E. James, Edward E. Waldrep, Douglas L. Delahanty, and Roman Cieslak. 2015. Trauma Coping Self-Efficacy: A context specific self-efficacy measure for traumatic stress. *Psychological Trauma: Theory, Research, Practice, and Policy* 7, 6 (2015), 591.

[17] Hugh Beyer, and Karen Holtzblatt. 1997. Contextual design: Defining customer-centered systems. In *Principles of Contextual Inquiry (Chapter 3)*. Elsevier, 41–66.

[18] Vincent Bindschaedler, Paul Grubbs, David Cash, Thomas Ristenpart, and Vitaly Shmatikov. 2018. The tao of inference in privacy-protected databases. *Proceedings of the VLDB Endowment* 11, 11 (2018), 1715–1728. DOI: https://doi.org/10.14778/3236187.3236217

[19] Sarah Bird, Krishnaram Kenthapadi, Emre Kiciman, and Margaret Mitchell. 2019. Fairness-aware machine learning: Practical challenges and lessons learned. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM'19).* ACM, 834–835. DOI: https://doi.org/10.1145/3289600.3291383

[20] Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan. 2017. Signal processing and machine learning for mental health research and clinical applications [perspectives]. *IEEE Signal Processing Magazine* 34, 5 (2017), 196–195. DOI: https://doi.org/10.1109/MSP.2017.2718581

[21] Katia Bourahmoune and Toshiyuki Amagasa. 2019. AI-powered posture training: Application of machine learning in sitting posture recognition using the lifechair smart cushion. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19).* Sarit Kraus (Ed.), AAAI Press, 5808–5814.

[22] Leo Breiman and Adele Cutler. 2007. *Random Forests-Classification Description.* Department of Statistics, University of California, Berkeley.

[23] Egon L. Broek, Frans Sluis, and Ton Dijkstra. 2013. Cross-validation of bimodal health-related stress assessment. *Personal Ubiquitous Computing* 17, 2 (2013), 215–227. DOI: http://dx.doi.org/10.1007/s00779-011-0468-z

[24] Joy Buolamwini, and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency (FAT'18).* 77–91.

[25] Danilo Bzdok, and Andreas Meyer-Lindenberg. 2018. Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 3, 3 (2018), 223–230. DOI: https://doi.org/10.1016/j.bpsc.2017.11.007

[26] Longbing Cao, Philip S. Yu, Chengqi Zhang, and Huaifeng Zhang. 2008. *Data Mining for Business Applications* (1st ed.). Springer Publishing Company, Incorporated.

[27] Bokai Cao, Lei Zheng, Chenwei Zhang, Philip S. Yu, Andrea Piscitello, John Zulueta, Olu Ajilore, Kelly Ryan, and Alex D. Leow. 2017. DeepMood: Modeling mobile phone typing dynamics for mood detection. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17).* ACM, 747–755. DOI: https://doi.org/10.1145/3097983.3098086

[28] Stevie Chancellor. 2018. Computational methods to understand deviant mental wellness communities. In *Proceedings of the Extended Abstracts CHI 2018.* ACM, Paper DC05. https://doi.org/10.1145/3170427.3173021

[29] Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT'19).* ACM, 79–88. DOI: https://doi.org/10.1145/3287560.3287587

[30] Stevie Chancellor, Eric P. S. Baumer, and Munmun De Choudhury. 2019. Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 147 (November 2019), 32 pages. DOI: https://doi.org/10.1145/3359249

[31] Keng-hao Chang, Matthew K. Chan, and John Canny. 2011. AnalyzeThis: Unobtrusive mental health monitoring by voice. In *Proceedings of the CHI'11 Extended Abstracts on Human Factors in Computing Systems (CHI EA'11).* ACM, 1951–1956. DOI: https://doi.org/10.1145/1979742.1979859

[32] Ritu Chauhan, and Harleen Kaur. 2017. A feature-based selection technique for reduction of large scale data. *International Journal of Data Analysis Techniques and Strategies* 9, 3 (2017), 207–221. DOI: https://doi.org/10.1504/IJDATS.2017.086630

[33] Xuetong Chen, Martin D. Sykora, Thomas W. Jackson, and Suzanne Elayan. 2018. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Proceedings of the Companion the Web Conference 2018 (WWW'18).* 1653–1660. DOI: https://doi.org/10.1145/3184558.3191624

[34] Prerna Chikersal, Danielle Belgrave, Gavin Doherty, Angel Enrique, Jorge E. Palacios, Derek Richards, and Anja Thieme. 2020. Understanding client support strategies to improve clinical outcomes in an online mental health intervention. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI'20).* ACM. DOI: http://dx.doi.org/10.1145/3313831.3376341

[35] James F. Childress, Ruth R. Faden, Ruth D. Gaare, Lawrence O. Gostin, Jeffrey Kahn, Richard J. Bonnie, Nancy E. Kass, Anna C. Mastroianni, Jonathan D. Moreno, and Phillip Nieburg. 2002. Public health ethics: Mapping the terrain. *The Journal of Law, Medicine & Ethics* 30, 2 (2002), 170–178. DOI: https://doi.org/10.1111/j.1748-720X.2002.tb00384.x

[36] Pietro Cipresso, Silvia Serino, Yuri Ostrovsky, and Justin T. Baker. 2018. Pervasive computing paradigms for mental health. In *Proceedings of the 7th International Conference, Mindcare 2018* (1st ed.). Springer Publishing Company, Incorporated.

[37] Moustapha M. Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. 2017. Houdini: Fooling deep structured visual and speech recognition mossdels with adversarial examples. In *Proceedings of the s Neural Information Processing Systems (NIPS'17).* 6977–6987.

[38] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A global measure of perceived stress. *Journal of Health and Social Behavior* 24, 4 (1983), 385–396. DOI: https://www.jstor.org/stable/2136404

[39] Jason A. Colquitt, and Jessica B. Rodell. 2015. Measuring justice and fairness. In *Oxford Handbook of Justice in the Workplace.* Vol. 187, Oxford University Press, 202.

[40] Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the 3rd Workshop on Computational Lingusitics and Clinical Psychology*, 106–117. DOI: https://www.aclweb.org/anthology/W16-0311

[41] David Coyle, Anja Thieme, Conor Linehan, Madeline Balaam, Jayne Wallace, and Siân Lindley. 2014. Emotional wellbeing. *International Journal of Human Computer Studies* 8, 72 (2014), 627–628. DOI: http://dx.doi.org/10.1016/j.ijhcs.2014.05.008

[42] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, 2098–2110. DOI: https://doi.org/10.1145/2858036.2858207

[43] Drew DeHaas, Jesse Craig, Colin Rickert, Margaret J. Eppstein, Paul Haake, and Kirsten Stor. 2007. Feature selection and classification in noisy epistatic problems using a hybrid evolutionary approach. In *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation (GECCO'07)*. ACM, 1872–1872. DOI: https://doi.org/10.1145/1276958.1277331

[44] Orianna DeMasi and Benjamin Recht. 2017. A step towards quantifying when an algorithm can and cannot predict an individual's wellbeing. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 763–771. DOI: https://doi.org/10.1145/3123024.3125609

[45] Joachim Diederich, Aqeel Al-Ajmi, and Peter Yellowlees. 2007. Ex-ray: Data mining and mental health. *Applied Soft Computing* 7, 3 (2007), 923–928. DOI: http://dx.doi.org/10.1016/j.asoc.2006.04.007

[46] Ed Diener, Derrick Wirtz, William Tov, Chu Kim-Prieto, Dong-won Choi, Shigehiro Oishi, and Robert Biswas-Diener. 2010. New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research* 97, 2 (2010), 143–156. DOI: https://doi.org/10.1007/s11205-009-9493-y

[47] Loretta Dipietro, Carl J. Caspersen, Adrian M. Ostfeld, and Ethan R. Nadel. 1993. A survey for assessing physical activity among older adults. *Medicine & Science in Sports & Exercise* 25, 5 (1993), 628–642. DOI: http://dx.doi.org/10.1249/00005768-199305000-00016

[48] Kevin Doherty, José Marcano-Belisario, Martin Cohn, Nikolaos Mastellos, Cecily Morrison, Josip Car, and Gavin Doherty. 2019. Engagement with mental health screening on mobile devices: Results from an antenatal feasibility study. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM, Paper 186, 15 pages. DOI: https://doi.org/10.1145/3290605.3300416

[49] Gavin Doherty, David Coyle, and John Sharry. 2012. Engagement with online mental health interventions: An exploratory clinical study of a treatment for depression. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*. ACM, 1421–1430. DOI: https://doi.org/10.1145/2207676.2208602

[50] Afsaneh Doryab, Mads Frost, et al. 2015. Impact factor analysis: Combining prediction with parameter ranking to reveal the impact of behavior on health outcome. *Personal Ubiquitous Computing* 19, 2 (2015), 355–365. DOI: http://dx.doi.org/10.1007/s00779-014-0826-8

[51] Georgios Drakos. 2018. How to select the right evaluation metric for machine learning models: Part 1 regression metrics. *Towards Data Science*. Last retrieved 6th of July 2019 from https://towardsdatascience.com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0.

[52] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI'18)*. ACM, 211–223. DOI: https://doi.org/10.1145/3172944.3172961

[53] Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM, Paper 134, 16 pages. DOI: https://doi.org/10.1145/3290605.3300364

[54] Emad N. Eskandar and Barry J. Richmond. 1991. Decoding of neuronal signals in visual pattern recognition. In *Proceedings of the 4th International Conference on Neural Information Processing Systems (NIPS'91)*. J. E. Moody, S. J. Hanson, and R. P. Lippmann (Eds.), Morgan Kaufmann Publishers Inc., San Francisco, CA, 356–363.

[55] Deborah Estrin, J. P. Pollak, and Tauhidur Rahman. 2017. MOBISYS MobiSys: Mobile systems, applications, and services. In *Proceedings of the 1st Workshop on Digital Biomarkers*. ACM. https://dl.acm.org/citation.cfm?id=3089341.

[56] Harris A. Eyre, Ajeet B. Singh, and Charles Reynolds III. 2016. Tech giants enter mental health. *World Psychiatry* 15, 1 (2016), 21–22. DOI: https://dx.doi.org/10.1002%2Fwps.20297

[57] Iram Fatima, Hamid Mukhtar, Hafiz Farooq Ahmad, and Kashif Rajpoot. 2018. Analysis of user-generated content from online social communities to characterise and predict depression degree. *Journal of Information Science* 44, 5 (October 2018), 683–695. DOI: https://doi.org/10.1177/0165551517740835

[58] Casey Fiesler, and Nicholas Proferes. 2018. "Participant" perceptions of Twitter research ethics. *Social Media + Society* 4, 1 (2018), 1–14. DOI: https://doi.org/10.1177%2F2056305118763366

[59] Thomas Filk and Albrecht von Müller. 2008. Evolutionary learning of small networks. *Complex* 13, 3 (2008), 43–54. DOI: http://dx.doi.org/10.1002/cplx.v13:3

[60] David Feil-Seifer and Maja J. Matarić. 2012. Distance-based computational models for facilitating robot Interaction with children. *Journal of Human-Robot Interaction* 1, 1 (2012), 55–77. https://doi.org/10.5898/JHRI.1.1.Feil-Seifer.

[61] Chaonan Feng, Huimin Gao, Xuefeng B. Ling, Jun Ji, and Yantao Ma. 2018. Shorten bipolarity checklist for the differentiation of subtypes of bipolar disorder using machine learning. In *Proceedings of the 2018 6th International Conference on Bioinformatics and Computational Biology (ICBCB'18)*. ACM, 162–166. DOI: https://doi.org/10.1145/3194480.3194508

[62] Joakim Ihle Frogner, Farzan Majeed Noori, Pål Halvorsen, Steven Alexander Hicks, Enrique Garcia-Ceja, Jim Torresen, and Michael Alexander Riegler. 2019. One-dimensional convolutional neural networks on motor activity measurements in detection of depression. In *Proceedings of the 4th International Workshop on Multimedia for Personal Health & Health Care (HealthMedia'19)*. ACM, 9–15. DOI: https://doi.org/10.1145/3347444.3356238

[63] Dimitrios Galiatsatos, Georgia Konstantopoulou, George Anastassopoulos, Marina Nerantzaki, Konstantinos Assimakopoulos, and Dimitrios Lymberopoulos. 2015. Classification of the most significant psychological symptoms in mental patients with depression using bayesian network. In *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS) (EANN'15)*. Lazaros Iliadis and Chrisina Jane (Eds.), ACM, Paper 15, 8 pages. DOI: https://doi.org/10.1145/2797143.2797159

[64] Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman Pathak. 2018. "Let me tell you about your mental health!": Contextualized classification of reddit posts to DSM-5 for web-based intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM'18)*. ACM, 753–762. DOI: https://doi.org/10.1145/3269206.3271732

[65] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L. Beam, and Rajesh Ranganath. 2018. Opportunities in machine learning for healthcare. arXiv:1806.00388. DOI: https://arxiv.org/abs/1806.00388

[66] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. arXiv:1803.09010. DOI: https://arxiv.org/abs/1803.09010

[67] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. 2016. Continuous stress detection using a wrist device: In laboratory and real life. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: AdjunctAdjunct (UbiComp'16)*. 1185–1193. DOI: https://doi.org/10.1145/2968219.2968306

[68] Russell E. Glasgow, Cynthia Vinson, David Chambers, Muin J. Khoury, Robert M. Kaplan, and Christine Hunter. 2012. National Institutes of Health approaches to dissemination and implementation science: Current and future directions. *American Journal of Public Health* 102, 7 (2012), 1274–1281. DOI: https://doi.org/10.2105/AJPH.2012.300755

[69] Bart N. Green, Claire D. Johnson, and Alan Adams. 2006. Writing narrative literature reviews for peer-reviewed journals: Secrets of the trade. *Journal of Chiropractic Medicine* 5, 3 (2006), 101–117. DOI: https://dx.doi.org/10.1016%2FS0899-3467(07)60142-6

[70] Ben Greer, Dan Robotham, Sara Simblett, Hannah Curtis, Helena Griffiths, and Til Wykes. 2019. Digital exclusion among mental health service users: Qualitative investigation. *Journal of Medical Internet Research* 21, 1 (2019), e11696. DOI: https://doi.org/10.2196/11696

[71] Thomas L. Griffiths, and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences* 101, 1 (2004), 5228–5235. DOI: https://doi.org/10.1073/pnas.0307752101

[72] Tayfun Gürel, Luc De Raedt, and Stefan Rotter. 2007. Ranking neurons for mining structure-activity relations in biological neural networks: NeuronRank. *Neurocomputing* 70, 10–12 (2007), 1897–1901. DOI: http://dx.doi.org/10.1016/j.neucom.2006.10.064

[73] Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. 2020. Toward fairness in AI for people with disabilities: A research roadmap. *SIGACCESS Access. Comput.* 125, Article 2 (October 2019), 1 page. DOI: https://doi.org/10.1145/3386296.3386298

[74] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. 2017. Countering adversarial images using input transformations. arXiv:1711.00117. DOI: https://arxiv.org/abs/1711.00117

[75] Kiwan Han, Jeonghun Ku, Kwanguk Kim, Hee Jeong Jang, Junyoung Park, Jae-Jin Kim, Chan Hyung Kim, Min-Hyung Choi, In Young Kim, and Sun I. Kim. 2009. Virtual reality prototype for measurement of expression characteristics in emotional situations. *Computers in Biology and Medicine* 39, 2 (2009), 173–179. http://dx.doi.org/10.1016/j.compbiomed.2008.12.002

[76] Hoda Heidari, Claudio Ferrari, Krishna P. Gummadi, and Andreas Krause. 2018. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. arXiv:1806.04959. DOI: https://arxiv.org/abs/1806.04959

[77] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. 2017. Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS'17)*. ACM, 95–99. DOI: https://doi.org/10.1145/3064663.3064703

[78] Tad Hirsch, Christina Soma, Kritzia Merced, Patty Kuo, Aaron Dembe, Derek D. Caperton, David C. Atkins, and Zac E. Imel. 2018. "It's hard to argue with a computer": Investigating Psychotherapists' Attitudes towards Automated Evaluation. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS'18)*. ACM, 559–571. DOI: https://doi.org/10.1145/3196709.3196776

[79] Kristina Höök. 2000. Steps to take before intelligent user interfaces become real. *Interacting with Computers* 12, 4 (2000), 409–426. http://dx.doi.org/10.1016/S0953-5438(99)00006-5

[80] Ayanna Howard, Cha Zhang, and Eric Horvitz. 2017. Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. In *Proceedings of the IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO'17)*. IEEE, 1–7. DOI: https://doi.org/10.1109/ARSO.2017.8025197

[81] Vignesh Jagadeesh, S. Karthikeyan, and B. S. Manjunath. 2010. Spatio-temporal optical flow statistics (STOFS) for activity classification. In *Proceedings of the 7th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'10)*. ACM, 178–182. DOI: http://dx.doi.org/10.1145/1924559.1924583

[82] Vidhi Jain and Prakhar Agarwal. 2017. Symptomatic diagnosis and prognosis of psychiatric disorders through personal gadgets. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA'17)*. ACM, 118–123. DOI: https://doi.org/10.1145/3027063.3048417

[83] Natasha Jaques, Sara Taylor, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2016. Multitask learning for predicting health, stress, and happiness. In *Proceedings of the NIPS Workshop on Machine Learning for Healthcare*. Last Retrieved 14th September 2019 from https://pdfs.semanticscholar.org/b228/7a406985980515d5cc63e9b37fb17c5186f8.pdf.

[84] Sooyeon Jeong and Cynthia Lynn Breazeal. 2016. Improving smartphone users' affect and wellbeing with personalized positive psychology interventions. *In Proceedings of the 4th International Conference on Human Agent Interaction (HAI'16)*. ACM, New York, NY, 131–137. DOI: https://doi.org/10.1145/2974804.2974831

[85] Michael I. Jordan, and Tom M. Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science* 349, 6245 (2015), 255–260. DOI: https://doi.org/10.1126/science.aaa8415

[86] Deepali J. Joshi, Mohit Makhija, Yash Nabar, Ninad Nehete, and Manasi S. Patwardhan. 2018. Mental health analysis using deep learning for feature extraction. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (CoDS-COMAD'18)*. 356–359. DOI: https://doi.org/10.1145/3152494.3167990

[87] K. A. Kasmiran, A. Y. Zomaya, A. A. Mazari, and R. J. Garsia. 2010. SVM-enabled prognostic method for clinical decision making: The use of CD4 T-cell level and HIV-1 viral load for guiding treatment initiation and alteration. In *Proceedings of the 2010 IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS'10)*. IEEE Computer Society, 19–25. DOI: http://dx.doi.org/10.1109/CBMS.2010.6042686

[88] Pavleen Kaur, Ravinder Kumar, and Munish Kumar. 2019. A healthcare monitoring system using random forest and internet of things (IoT). *Multimedia Tools Applications* 78, 14 (July 2019), 19905–19916. DOI: https://doi.org/10.1007/s11042-019-7327-8

[89] Ramakanth Kavuluru, María Ramos-Morales, Tara Holaday, Amanda G. Williams, Laura Haye, and Julie Cerel. 2016. Classification of Helpful Comments on Online Suicide Watch Forums. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB'16)*. ACM, 32–40. DOI: https://doi.org/10.1145/2975167.2975170

[90] Ronald C. Kessler et al. 2007. Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry* 6, 3 (2007), 168. DOI: s

[91] Ronald C. Kessler et al. 2008. Individual and societal effects of mental disorders on earnings in the United States: Results from the national comorbidity survey replication. *American Journal of Psychiatry* 165, 6 (2008), 703–711. DOI: https://doi.org/10.1176/appi.ajp.2008.08010126

[92] Alex V. Kotlar and Thomas S. Wingo. 2018. Tutorial: Rapidly Identifying Disease-associated Rare Variants using Annotation and Machine Learning at Whole-genome Scale Online. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB'18)*. ACM, 558–558. DOI: https://doi.org/10.1145/3233547.3233666

[93] John R. Koza, Forrest H. Bennett, David Andre, and Martin A. Keane. 1996. Automated design of both the topology and sizing of analog electrical circuits using genetic programming. In *Proceedings of the Artificial Intelligence in Design'96*. Springer, 151–170. DOI: https://doi.org/10.1007/978-94-009-0279-4_9

[94] Jens Kremkow, Arvind Kumar, Stefan Rotter, and Ad Aertsen. 2007. Emergence of population synchrony in a layered network of the cat visual cortex. *Neurocomputing* 70, 10–12 (2007), 2069–2073. DOI: http://dx.doi.org/10.1016/j.neucom.2006.10.130

[95] Nikolaus Kriegeskorte, Jerzy Bodurka, and Peter Bandettini. 2008. Artifactual time-course correlations in echoplanar fMRI with implications for studies of brain function. *International Journal of Imaging Systems and Technology* 18, 5–6 (2008), 345–349. DOI: http://dx.doi.org/10.1002/ima.v18:5/6

[96] Birgit Kriener, Tom Tetzlaff, Ad Aertsen, Markus Diesmann, and Stefan Rotter. 2008. Correlations and population dynamics in cortical networks. *Neural Computing* 20, 9 (2008), 2185–2226. DOI: http://dx.doi.org/10.1162/neco.2008.02-07-474

[97] Kurt Kroenke, and Robert L. Spitzer. 2002. The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals* 32, 9 (2002), 509–515. DOI: https://doi.org/10.3928/0048-5713-20020901-06

[98] Arvind Kumar, Sven Schrader, Ad Aertsen, and Stefan Rotter. 2008. The high-conductance state of cortical networks. *Neural Computing* 20, 1 (2008), 1–43. DOI: http://dx.doi.org/10.1162/neco.2008.20.1.1

[99] Reeva Lederman, John Gleeson, Greg Wadley, Simon D'alfonso, Simon Rice, Olga Santesteban-Echarri, and Mario Alvarez-Jimenez. 2019. Support for carers of young people with mental illness: Design and trial of a technology-mediated therapy. *ACM Transactions on Computer-Human Interaction* 26, 1, Article 4 (2019), 33 pages. DOI: https://doi.org/10.1145/3301421

[100] Yena Lee, Renee-Marie Ragguett, Rodrigo B. Mansur, Justin J. Boutilier, Joshua D. Rosenblat, Alisson Trevizol, Elisa Brietzke, et al. 2018. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders* 241, 519–532. DOI: https://doi.org/10.1016/j.jad.2018.08.073

[101] Yugyung Lee, Saranya Krishnamoorthy, and Deendayal Dinakarpandian. 2013. A semantic framework for intelligent matchmaking for clinical trial eligibility criteria. *ACM Transactions on Intelligent Systems and Technology* 4, 4, Article 71 (2013), 32 pages. DOI: https://doi.org/10.1145/2508037.2508052

[102] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 182 (2019), 26 pages. DOI: https://doi.org/10.1145/3359284

[103] Sidney R. Lehky. 2004. Bayesian estimation of stimulus responses in Poisson spike trains. *Neural Computing* 16, 7 (2004), 1325–1343. DOI: http://dx.doi.org/10.1162/089976604323057407

[104] Alessandro Liberati, Douglas G. Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C. Gøtzsche, John P. A. Ioannidis, Mike Clarke, Pl. J. Devereaux, Jos Kleijnen, and David Moher. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PloS Medicine* 6, 7 (2009), e1000100. DOI: https://doi.org/10.1371/journal.pmed.1000100

[105] Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Qi Li, Jie Huang, Lianhong Cai, and Ling Feng. 2014. User-level psychological stress detection from social media using deep neural network. In *Proceedings of the 22nd ACM international conference on Multimedia (MM'14)*. ACM, 507–516. DOI: https://doi.org/10.1145/2647868.2654945

[106] Zachary C. Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (May-June 2018), 31–57. DOI: https://doi.org/10.1145/3236386.3241340

[107] Fannie Liu. 2019. Expressive biosignals: Authentic social cues for social connection. In *Proceedings of the Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA'19)*. ACM, Paper DC12, 5 pages. DOI: https://doi.org/10.1145/3290607.3299081

[108] Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics* 75, S (2017), S34–S42. DOI: https://doi.org/10.1016/j.jbi.2017.05.023

[109] Joseph B. Lyons, Garrett G. Sadler, Kolina Koltai, Henri Battiste, Nhut T. Ho, Lauren C. Hoffmann, David Smith, Walter Johnson, and Robert Shively. 2017. Shaping trust through transparent design: Theoretical and experimental guidelines. In *Proceedings of the Advances in Human Factors in Robots and Unmanned Systems*. Springer, 127–136. DOI: https://doi.org/10.1007/978-3-319-41959-6_11

[110] Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland. 2010. Social sensing for epidemiological behavior change. In *Proceedings of the 12th ACM international conference on Ubiquitous computing (UbiComp'10)*. ACM, 291–300. DOI: https://doi.org/10.1145/1864349.1864394

[111] Maria Madsen, and Shirley Gregor. 2000. Measuring human-computer trust. In *Proceedings of the 11th Australasian Conference on Information Systems*. Vol. 53, 6–8.

[112] Adria Mallol-Ragolta, Svati Dhamija, and Terrance E. Boult. 2018. A multimodal approach for predicting changes in PTSD symptom severity. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI'18)*. ACM, 324–333. DOI: https://doi.org/10.1145/3242969.3242981

[113] Mirza Mansoor Baig, Hamid Gholamhosseini, Aasia A. Moqeem, Farhaan Mirza, and Maria Lindén. 2017. A systematic review of wearable patient monitoring systems—Current challenges and opportunities for clinical adoption. *Journal of Medical Systems* 41, 7 (2017), 1–9. DOI: https://doi.org/10.1007/s10916-017-0760-1

[114] Martin Maritsch, Caterina Bérubé, Mathias Kraus, Vera Lehmann, Thomas Züger, Stefan Feuerriegel, Tobias Kowatsch, and Felix Wortmann. 2019. Improving heart rate variability measurements from consumer smartwatches with machine learning. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers (UbiComp/ISWC'19 Adjunct)*. ACM, 934–938. DOI: https://doi.org/10.1145/3341162.3346276

[115] Maja J. Matarić. 2019. Human-machine and human-robot interaction for long-term user engagement and behavior change. In *Proceedings of the 25th Annual International Conference on Mobile Computing and Networking (MobiCom'19)*. ACM, Paper 56, 2 pages. DOI:https://doi.org/10.1145/3300061.3300141

[116] Andrew M. McIntosh, Robert Stewart, Ann John, Daniel J. Smith, Katrina Davis, Cathie Sudlow, Aiden Corvin, et al. 2016. Data science for mental health: A UK perspective on a global challenge. *The Lancet Psychiatry* 3, 10 (2016), 993–998. DOI:https://doi.org/10.1016/S2215-0366(16)30089-X

[117] Quinten McNamara, Alejandro De La Vega, and Tal Yarkoni. 2017. Developing a comprehensive framework for multimodal feature extraction. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. ACM, 1567–1574. DOI:https://doi.org/10.1145/3097983.3098075

[118] Abhinav Mehrotra and Mirco Musolesi. 2017. Designing effective movement digital biomarkers for unobtrusive emotional state mobile monitoring. In *Proceedings of the 1st Workshop on Digital Biomarkers (DigitalBiomarkers'17)*. ACM, 3–8. DOI:https://doi.org/10.1145/3089341.3089342

[119] Susan Michie, Lucy Yardley, Robert West, Kevin Patrick, and Felix Greaves. 2017. Developing and evaluating digital interventions to promote behavior change in health and health care: Recommendations resulting from an international workshop. *Journal of Medical Internet Research* 19, 6 (2017), e232. DOI:https://doi.org/10.2196/jmir.7126

[120] Gatis Mikelsons, Abhinav Mehrotra, Mirco Musolesi, and Nigel Shadbolt. 2019. Evaluating Machine Learning Algorithms for Prediction of the Adverse Valence Index Based on the Photographic Affect Meter. In *Proceedings of the 5th ACM Workshop on Mobile Systems for Computational Social Science (MCSS'19)*. ACM, New York, NY, 5–10. DOI:https://doi.org/10.1145/3325426.3329948

[121] William R. Miller, Theresa B. Moyers, Denise Ernst, and Paul Amrhein. 2003. *Manual for the Motivational Interviewing Skill Code (MISC), Version 2.1*. Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico. Last retrieved 14th of September from https://casaa.unm.edu/download/misc.pdf.

[122] Vikramjit Mitra, Elizabeth Shriberg, Mitchell McLaren, Andreas Kathol, Colleen Richey, Dimitra Vergyri, and Martin Graciarena. 2014. The SRI AVEC-2014 evaluation system. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC'14)*. ACM, 93–101. DOI:https://doi.org/10.1145/2661806.2661818

[123] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G. Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine* 151, 4 (2009), 264–269. DOI:https://doi.org/10.7326/0003-4819-151-4-200908180-00135

[124] David C. Mohr, Mi Zhang, and Stephen M. Schueller. 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology* 13, 23–47. DOI:https://doi.org/10.1146/annurev-clinpsy-032816-044949

[125] David C. Mohr, Kathryn Noth Tomasino, Emily G. Lattie, Hannah L. Palac, Mary J. Kwasny, Kenneth Weingardt, Chris J. Karr, et al. 2017. IntelliCare: An eclectic, skills-based app suite for the treatment of depression and anxiety. *Journal of Medical Internet Research* 19, 1 (2017), e10. DOI:https://doi.org/10.2196/jmir.6645

[126] Stuart A. Montgomery, and M. A. R. I. E. Åsberg. 1979. A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry* 134, 4 (1979), 382–389. DOI:https://doi.org/10.1192/bjp.134.4.382

[127] Cecily Morrison, and Gavin Doherty. 2014. Analyzing engagement in a web-based intervention platform through visualizing log-data. *Journal of Medical Internet Research* 16, 11 (2014), e252. DOI:https://doi.org/10.2196/jmir.3575

[128] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K. D'Mello, Munmun De Choudhury, Gregory D. Abowd, and Thomas Plötz. 2019. Prediction of mood instability with passive sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3, Article 75 (2019), 21 pages. DOI:https://doi.org/10.1145/3351233

[129] T. B. Moyers, T. Martin, J. K. Manuel, W. R. Miller, and D. Ernst. 2010. *Revised global scales: Motivational interviewing treatment integrity 3.1. 1* (MITI 3.1. 1). Unpublished manuscript, University of New Mexico, Albuquerque, NM.

[130] Inbal Nahum-Shani, Shawna N. Smith, Bonnie J. Spring, Linda M. Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A. Murphy. 2017. Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* 52, 6 (2017), 446–462. DOI:https://doi.org/10.1007/s12160-016-9830-8

[131] Hiroyuki Nakahara, Shun-Ichi Amari, and Barry J. Richmond. 2006. A Comparison of Descriptive Models of a Single Spike Train by Information-Geometric Measure. *Neural Computing* 18, 3 (2006), 545–568. DOI:http://dx.doi.org/10.1162/089976606775623289

[132] Lisa P. Nathan, Anja Thieme, Deborah Tatar, and Stacy Branham. 2017. Disruptions, dilemmas and paradoxes: Ethical matter(s) in design research. *Interacting with Computers* 29, 1 (2017), 1–9. DOI:https://doi.org/10.1093/iwc/iww034

[133] Thin Nguyen, Bridianne O'Dea, Mark Larsen, Dinh Phung, Svetha Venkatesh, and Helen Christensen. 2017. Using linguistic and topic analysis to classify sub-groups of online depression communities. *Multimedia Tools and Applications* 76, 8 (2017), 10653–10676. DOI:https://doi.org/10.1007/s11042-015-3128-x

[134] Alicia L. Nobles, Jeffrey J. Glenn, Kamran Kowsari, Bethany A. Teachman, and Laura E. Barnes. 2018. Identification of imminent suicide risk among young adults using text messages. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. ACM, Paper 413, 11 pages. DOI:https://doi.org/10.1145/3173574.3173987

[135] Ehimwenma Nosakhare and Rosalind Picard. 2019. Probabilistic latent variable modeling for assessing behavioral influences on well-being. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19)*. ACM, 2718–2726. DOI: https://doi.org/10.1145/3292500.3330738

[136] Blessing Ojeme, and Audrey Mbogho. 2016. Selecting learning algorithms for simultaneous identification of depression and comorbid disorders. *Procedia Computer Science* 96, 1294–1303. DOI: https://doi.org/10.1016/j.procs.2016.08.174

[137] Kathleen O'Leary, Stephen M. Schueller, Jacob O. Wobbrock, and Wanda Pratt. 2018. "Suddenly, we got to become therapists for each other": Designing peer support chats for mental health. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. ACM, Paper 331, 14 pages. DOI: https://doi.org/10.1145/3173574.3173905

[138] Sharon Oviatt, Björn Schuller, Philip R. Cohen, Daniel Sonntag, Gerasimos Potamianos, and Antonio Krüger (Eds.). 2018. *The Handbook of Multimodal-Multisensor Interfaces*. ACM and Morgan & Claypool, xvii–xix. DOI: https://doi.org/10.1145/3107990.3107991

[139] Theodor Chris Panagiotakopoulos, Dimitrios Panagiotis Lyras, Miltos Livaditis, Kyriakos N. Sgarbas, George C. Anastassopoulos, and Dimitrios K. Lymberopoulos. 2010. A contextual data mining approach toward assisting the treatment of anxiety disorders. *IEEE Transactions on Information Technology in Biomedicine* 14, 3 (2010), 567–581. DOI: https://doi.org/10.1109/TITB.2009.2038905

[140] Pablo Paredes, Ran Gilad-Bachrach, Mary Czerwinski, Asta Roseway, Kael Rowan, and Javier Hernandez. 2014. PopTherapy: Coping with stress through pop-culture. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth'14)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 109–117. DOI: http://dx.doi.org/10.4108/icst.pervasivehealth.2014.255070

[141] Albert Park, Mike Conway, and Annie T. Chen. 2018. Examining thematic similarity, difference, and membership in three online mental health communities from Reddit: A text mining and visualization approach. *Computers in Human Behavior* 78, 98–112. DOI: https://doi.org/10.1016/j.chb.2017.09.001

[142] Ives Cavalcante Passos, Benson Mwangi, and Flavio Kapczinski. 2019. *Personalized Psychiatry: Big Data Analytics in Mental Health* (1st ed.). Springer Publishing Company, Incorporated.

[143] Mensah Kwabena Patrick. 2015. Textual prediction of attitudes towards mental health. *International Journal of Knowledge Engineering and Data Mining* 3, 3/4 (2015), 274–285. DOI: http://dx.doi.org/10.1504/IJKEDM.2015.074076

[144] David A. Patterson and Richard N. Cloud. 2000. The application of artificial neural networks for outcome prediction in a cohort of severely mentally ill outpatients. *Journal of Technology and Human Services* 16, 2/3 (2000), 47–61. DOI: http://dx.doi.org/10.1300/J017v16n02_05

[145] John P. Pestian, Pawel Matykiewicz, and Jacqueline Grupp-Phelan. 2008. Using natural language processing to classify suicide notes. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP'08)*. ACM, Stroudsburg, PA, 96–97.

[146] Lawrence Pfeffer, David Ide, Craig Stewart, and Dietmar Plenz. 2004. A life support system for stimulation of and recording from rodent neuron networks grown on multi-electrode arrays. In *Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems (CBMS'04)*. IEEE Computer Society, 473–478. DOI: http://dx.doi.org/10.1109/CBMS.2004.5

[147] André Pimenta, Sergio Gonçalves, Davide Carneiro, Florentino Fde-riverola, José Neves, and Paulo Novais. 2015. Mental workload management as a tool in e-learning scenarios. In *Proceedings of the 5th International Conference on Pervasive and Embedded Computing and Communication Systems (PECCS'15)*. César Benavente-Peces, Olivier Paillet, and Andreas Ahrens (Eds.), SCITEPRESS—Science and Technology Publications, 25–32. DOI: https://doi.org/10.5220/0005237700250032

[148] John P. Pollak, Phil Adams, and Geri Gay. 2011. PAM: A photographic affect meter for frequent, in situ measurement of affect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*. ACM, 725–734. DOI: https://doi.org/10.1145/1978942.1979047

[149] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. arXiv:1802.07810. DOI: https://arxiv.org/abs/1802.07810

[150] Yada Pruksachatkun, Sachin R. Pendse, and Amit Sharma. 2019. Moments of change: Analyzing peer-based cognitive support in online mental health forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM, Paper 64, 1–13. DOI: https://doi.org/10.1145/3290605.3300294

[151] Alessandro Puiatti, Steven Mudda, Silvia Giordano, and Oscar Mayora. 2011. Smartphone-centred wearable sensors network for monitoring patients with bipolar disorder. In *Proceedings of the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 3644–3647. DOI: https://doi.org/10.1109/IEMBS.2011.6090613

[152] Thomas Quisel, Wei-Nchih Lee, and Luca Foschini. 2017. Observation time vs. performance in digital phenotyping. In *Proceedings of the 1st Workshop on Digital Biomarkers (DigitalBiomarkers'17)*. ACM, 33–36. DOI: https://doi.org/10.1145/3089341.3089347

[153] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. 2011. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of the 13th international conference on Ubiquitous computing (UbiComp'11)*. ACM, 385–394. DOI: https://doi.org/10.1145/2030112.2030164

[154] Neelesh Rastogi, Fazel Keshtkar, and Md Suruz Miah. 2018. A multi-modal human robot interaction framework based on cognitive behavioral therapy model. In *Proceedings of the Workshop on Human-Habitat for Health (H3): Human-Habitat Multimodal Interaction for Promoting Health and Well-Being in the Internet of Things Era (H3'18)*. ACM, Paper 2, 6 pages. DOI: https://doi.org/10.1145/3279963.3279968

[155] Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. 2019. Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC'19)*. ACM, 81–88. DOI: https://doi.org/10.1145/3347320.3357697

[156] Stefan Rennick-Egglestone, Sarah Knowles, Gill Toms, Penny Bee, Karina Lovell, and Peter Bower. 2016. Health technologies sin the wild': Experiences of engagement with computerised CBT. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, 2124–2135. DOI: https://doi.org/10.1145/2858036.2858128

[157] Derek Richards, Ladislav Timulak, Emma O'Brien, Claire Hayes, Noemi Vigano, John Sharry, and Gavin Doherty. 2015. A randomized controlled trial of an internet-delivered treatment: Its potential as a low-intensity community intervention for adults with symptoms of depression. *Behaviour Research and Therapy* 75, 20–31. DOI: https://doi.org/10.1016/j.brat.2015.10.005

[158] Darius A. Rohani, Maria Faurholt-Jepsen, Lars Vedel Kessing, and Jakob E. Bardram. 2018. Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: Systematic review. *JMIR mHealth and uHealth* 6, 8 (2018), e165. DOI: https://doi.org/10.2196/mhealth.9691

[159] Anne K. Rosenstiel, and Francis J. Keefe. 1983. The use of coping strategies in chronic low back pain patients: Relationship to patient characteristics and current adjustment. *Pain* 17, 1 (1983), 33–44. DOI: https://doi.org/10.1016/0304-3959(83)90125-2

[160] Sushmita Roy, Terran Lane, and Margaret Werner-Washburne. 2009. Learning structurally consistent undirected probabilistic graphical models. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*. ACM, New York, NY, 905–912. DOI: http://dx.doi.org/10.1145/1553374.1553490

[161] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1 (2019), 206–215. DOI: https://doi.org/10.1038/s42256-019-0048-x

[162] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161. http://dx.doi.org/10.1037/h0077714

[163] Daniel W. Russell. 1996. UCLA loneliness scale (version 3): Reliability, validity, and factor structure. *Journal of Personality Assessment* 66, 1 (1996), 20–40. DOI: https://doi.org/10.1207/s15327752jpa6601_2

[164] Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C. Mohr, and Konrad P. Kording. 2016. Voodoo machine learning for clinical predictions. *Biorxiv*, 059774. DOI: https://doi.org/10.1101/059774

[165] Koustuv Saha and Munmun De Choudhury. 2017. Modeling stress with social media around incidents of gun violence on college campuses. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW, Article 92 (2017), 27 pages. DOI: https://doi.org/10.1145/3134727

[166] Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kıcıman, and Munmun De Choudhury. 2019. A social media study on the effects of psychiatric medication use. In *Proceedings of the International AAAI Conference on Web and Social Media* 13, 1 (2019), 440–451. DOI: https://wvvw.aaai.org/ojs/index.php/ICWSM/article/view/3242

[167] Koustuv Saha, Sang Chan Kim, Manikanta D. Reddy, Albert J. Carter, Eva Sharma, Oliver L. Haimson, and Munmun De Choudhury. 2019. The language of LGBTQ± minority stress experiences on social media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 89 (November 2019), 22 pages. DOI: https://doi.org/10.1145/3361108

[168] Asif Salekin, Jeremy W. Eberle, Jeffrey J. Glenn, Bethany A. Teachman, and John A. Stankovic. 2018. A weakly supervised learning framework for detecting social anxiety and depression. *Proceedings of the ACM on Interactive Mobile Wearable Ubiquitous Technology* 2, 2, Article 81 (2018), 26 pages. DOI: https://doi.org/10.1145/3214284

[169] SAMHS (Substance Abuse and Mental Health Services Administration). 2018. Key substance use and mental health indicators in the United States: Results from the 2017 National Survey on Drug Use and Health (HHS Publication No. SMA 18-5068, NSUDH Series H-53). Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, Rockville, MD. Retrieved from https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHFFR2017/NSDUHFFR2017.pdf.

[170] Pedro Sanches, Axel Janson, Pavel Karpashevich, Camille Nadal, Chengcheng Qu, Claudia Daudén Roquet, Muhammad Umair, Charles Windlin, Gavin Doherty, Kristina Höök, and Corina Sas. 2019. HCI and affective health: Taking stock of a decade of studies and charting future research directions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM, Paper 245, 17 pages. DOI: https://doi.org/10.1145/3290605.3300475

[171] Jessica Schroeder, Chelsey Wilkes, Kael Rowan, Arturo Toledo, Ann Paradiso, Mary Czerwinski, Gloria Mark, and Marsha M. Linehan. 2018. Pocket Skills: A conversational mobile web app to support dialectical behavioral therapy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. ACM, Paper 398, 15 pages. DOI: https://doi.org/10.1145/3173574.3173972

[172] Stephen M. Schueller, Adrian Aguilera, and David C. Mohr. 2017. Ecological momentary interventions for depression and anxiety. *Depression and Anxiety* 34, 6 (2017), 540–545. DOI: https://doi.org/10.1002/da.22649

[173] Adrian B. R. Shatte, Delyse M. Hutchinson, and Samantha J. Teague. 2019. Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine* 49, 9 (2019), 1426–1448. DOI: https://doi.org/10.1017/S0033291719000151

[174] Hakim Sidahmed, Elena Prokofyeva, and Matthew B. Blaschko. 2016. Discovering predictors of mental health service utilization with k-support regularized logistic regression. *Information Sciences* 329, C (2016), 937–949. DOI: https://doi.org/10.1016/j.ins.2015.03.069

[175] Insu Song, Denise Dillon, Tze Jui Goh, and Min Sung. 2011. A health social network recommender system. In *Proceedings of the 14th international conference on Agents in Principle, Agents in Practice (PRIMA'11)*. David Kinny, Jane Yung-jen Hsu, Guido Governatori, and Aditya K. Ghose (Eds.), Springer, 361–372. DOI: http://dx.doi.org/10.1007/978-3-642-25044-6_29

[176] Dimitris Spathis, Sandra Servia-Rodriguez, Katayoun Farrahi, Cecilia Mascolo, and Jason Rentfrow. 2019. Passive mobile sensing and psychological traits for large scale mood prediction. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth'19)*. ACM, 272–281. DOI: https://doi.org/10.1145/3329189.3329213

[177] Dimitris Spathis, Sandra Servia-Rodriguez, Katayoun Farrahi, Cecilia Mascolo, and Jason Rentfrow. 2019. Sequence multi-task learning to forecast mental wellbeing from sparse self-reported data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19)*. ACM, 2886–2894. DOI: https://doi.org/10.1145/3292500.3330730

[178] B. Sri Nandhini and J. I. Sheeba. 2015. cyberbullying detection and classification using information retrieval algorithm. In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering and Technology (ICARCSET'15)*. ACM, Paper 20, 5 pages. DOI: http://dx.doi.org/10.1145/2743065.2743085

[179] M. Srividya, S. Mohanavalli, and N. Bhalaji. 2018. Behavioral modeling for mental health using machine learning algorithms. *Journal of Medical Systems* 42, 5 (2018), 88. DOI: https://doi.org/10.1007/s10916-018-0934-5

[180] Benjamin Staude, Stefan Rotter, and Sonja Grün. 2008. Can spike coordination be differentiated from rate covariation?. *Neural Computing* 20, 8 (2008), 1973–1999. DOI: http://dx.doi.org/10.1162/neco.2008.06-07-550

[181] Klaas E. Stephan, Florian Schlagenhauf, Quentin J. M. Huys, Sudhir Raman, Eduardo A. Aponte, Kay Henning Brodersen, Lionel Rigoux, et al. 2017. Computational neuroimaging strategies for single patient predictions. *Neuroimage* 145, Pt. B (2017), 180–199. DOI: https://doi.org/10.1016/j.neuroimage.2016.06.038

[182] Shelley E. Taylor, William T. Welch, Heejung S. Kim, and David K. Sherman. 2007. Cultural differences in the impact of social support on psychological and biological stress responses. *Psychological Science* 18, 9 (2007), 831–837. DOI: https://doi.org/10.1111/j.1467-9280.2007.01987.x

[183] Teewoon Tan, Ling Guan, and John Burne. 1999. A real-time image analysis system for computer-assisted diagnosis of neurological disorders. *Real-Time Imaging* 5, 4 (1999), 253–269. DOI: http://dx.doi.org/10.1006/rtim.1998.0139

[184] Leili Tavabi. 2019. Multimodal machine learning for interactive mental health therapy. In *Proceedings of the 2019 International Conference on Multimodal Interaction (ICMI'19)*. Wen Gao, Helen Mei Ling Meng, Matthew Turk, Susan R. Fussell, Björn Schuller, Yale Song, and Kai Yu (Eds.), ACM, 453–456. DOI: https://doi.org/10.1145/3340555.3356095

[185] Tom Tetzlaff, Stefan Rotter, Eran Stark, Moshe Abeles, Ad Aertsen, and Markus Diesmann. 2008. Dependence of neuronal correlations on filter characteristics and marginal spike train statistics. *Neural Computing* 20, 9 (2008), 2133–2184. http://dx.doi.org/10.1162/neco.2008.05-07-525

[186] Oliver Theobald. 2017. *Machine Learning for Absolute Beginners. A Plain English Introduction.* Scatterplot Press.

[187] Anja Thieme, John McCarthy, Paula Johnson, Stephanie Phillips, Jayne Wallace, Siân Lindley, Karim Ladha, Daniel Jackson, Diana Nowacka, Ashur Rafiev, Cassim Ladha, Thomas Nappey, Mathew Kipling, Peter Wright, Thomas D. Meyer, and Patrick Olivier. 2016. Challenges for designing new technology for health and wellbeing in a complex mental healthcare context. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, 2136–2149. DOI: https://doi.org/10.1145/2858036.2858182

[188] Anja Thieme, Jayne Wallace, Thomas D. Meyer, and Patrick Olivier. 2015. Designing for mental wellbeing: Towards a more holistic approach in the treatment and prevention of mental illness. In *Proceedings of the 2015 British HCI Conference (British HCI'15)*. ACM, New York, NY, 1–10. DOI: http://dx.doi.org/10.1145/2783446.2783586

[189] Anja Thieme, Danielle Belgrave, Akane Sano, and Gavin Doherty. 2020. Reflections on mental health assessment and ethics for machine learning applications. *Interactions* 27, 2 (2020), 6–7. DOI: https://doi.org/10.1145/3381342

[190] Graham Thornicroft, Diana Rose, Aliya Kassam, and Norman Sartorius. 2007. Stigma: Ignorance, prejudice or discrimination? *The British Journal of Psychiatry* 190, 3 (2007), 192–193. DOI: https://doi.org/10.1192/bjp.bp.106.025791

[191] John Torous, and Camille Nebeker. 2017. Navigating ethics in the digital age: Introducing connected and open research ethics (CORE), a tool for researchers and institutional review boards. *Journal of Medical Internet Research* 19, 2 (2017), e38. DOI: https://doi.org/10.2196/jmir.6793

[192] Truyen Tran, Dinh Phung, Wei Luo, Richard Harvey, Michael Berk, and Svetha Venkatesh. 2013. An integrated framework for suicide risk prediction. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13)*. Rayid Ghani, Ted E. Senator, Paul Bradley, Rajesh Parekh, and Jingrui He (Eds.), ACM, 1410–1418. DOI: https://doi.org/10.1145/2487575.2488196

[193] Konstantinos Tsiakas, Lynette Watts, Cyril Lutterodt, Theodoros Giannakopoulos, Alexandros Papangelis, Robert Gatchel, Vangelis Karkaletsis, and Fillia Makedon. 2015. A multimodal adaptive dialogue manager for depressive and anxiety disorder screening: A wizard-of-oz experiment. In *Proceedings of the 8th ACM International Conference on Pervasive Technologies Related to Assistive Environments (PETRA'15)*. ACM, Paper 82, 4 pages. DOI: https://doi.org/10.1145/2769493.2769572

[194] Robin Turkington, Maurice Mulvenna, Raymond Bond, Siobhan O'Neill, and Cherie Armour. 2018. The application of user event log data for mental health and wellbeing analysis. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI'18)*. BCS Learning and Development Ltd., Swindon, UK, Paper 4, 14 pages. DOI: https://doi.org/10.14236/ewic/HCI2018.4

[195] Jessica A. Turner, K. C. Anderson, and R. M. Siegel. 2003. Cell responsiveness in macaque superior temporal polysensory area measured by temporal discriminants. *Neural Computing* 15, 9 (2003), 2067–2090. http://dx.doi.org/10.1162/089976603322297296

[196] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC'16)*. ACM, 3–10. DOI: https://doi.org/10.1145/2988257.2988258

[197] Jose Juan Dominguez Veiga and Tomas E. Ward. 2016. Data collection requirements for mobile connected health: An end user development approach. In *Proceedings of the 1st International Workshop on Mobile Development (Mobile! 2016)*. ACM, 23–30. DOI: https://doi.org/10.1145/3001854.3001856

[198] Mélodie Vidal, Andreas Bulling, and Hans Gellersen. 2011. Analysing EOG signal features for the discrimination of eye movements with wearable devices. In *Proceedings of the 1st International Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction (PETMEI'11)*. ACM, 15–20. DOI: http://dx.doi.org/10.1145/2029956.2029962

[199] Nicole Voges, Ad Aertsen, and Stefan Rotter. 2007. Statistical analysis of spatially embedded networks: From grid to random node positions. *Neurocomputing* 70, 10–12 (2007), 1833–1837. DOI: http://dx.doi.org/10.1016/j.neucom.2006.10.126

[200] Tuong Manh Vu, Charlotte Probst, Joshua M. Epstein, Alan Brennan, Mark Strong, and Robin C. Purshouse. 2019. Toward inverse generative social science using multi-objective genetic programming. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'19)*. Manuel López-Ibáñez (Ed.), ACM, 1356–1363. DOI: https://doi.org/10.1145/3321707.3321840

[201] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'14)*. ACM, 3–14. DOI: https://doi.org/10.1145/2632048.2632054

[202] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology* 54, 6 (1988), 1063.

[203] Frank W. Weathers, Brett T. Litz, Debra S. Herman, Jennifer A. Huska, and Terence M. Keane. 1993. The PTSD checklist (PCL): Reliability, validity, and diagnostic utility. In *Proceedings of the Annual Convention of the International Society for Traumatic Stress Studies*. Vol. 462.

[204] Harvey A. Whiteford, Louisa Degenhardt, Jürgen Rehm, Amanda J. Baxter, Alize J. Ferrari, Holly E. Erskine, Fiona J. Charlson, et al. 2013. Global burden of disease attributable to mental and substance use disorders: Findings from the Global Burden of Disease Study 2010. *The Lancet* 382, 9904 (2013), 1575–1586. DOI: https://doi.org/10.1016/S0140-6736(13)61611-6

[205] Paula Wilbourne, Geralyn Dexter, and David Shoup. 2018. Research driven: Sibly and the transformation of mental health and wellness. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth'18)*. ACM, 389–391. DOI: https://doi.org/10.1145/3240925.3240932

[206] Choong-Wan Woo, Luke J. Chang, Martin A. Lindquist, and Tor D. Wager. 2017. Building better biomarkers: Brain models in translational neuroimaging. *Nature Neuroscience* 20, 3 (2017), 365–377. DOI:https://doi.org/10.1038/nn.4478

[207] World Health Organization (WHO). 2018. Fact sheets: Depression. Last retrieved 11th June 2019 from http://www.who.int/mediacentre/factsheets/fs369/en/.

[208] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R. Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys* 50, 2, Article 25 (2017), 33 pages. DOI:https://doi.org/10.1145/3057270

[209] Hui Yang and Peter A. Bath. 2019. Automatic prediction of depression in older age. In *Proceedings of the 3rd International Conference on Medical and Health Informatics 2019 (ICMHI'19)*. ACM, 36–44. DOI:https://doi.org/10.1145/3340037.3340042

[210] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM, Paper 238, 11 pages. DOI:https://doi.org/10.1145/3290605.3300468

[211] Amir Hossein Yazdavar, Hussein S. Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (ASONAM'17)*. Jana Diesner, Elena Ferrari, and Guandong Xu (Eds.), ACM, 1191–1198. DOI:https://doi.org/10.1145/3110025.3123028

[212] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM, Paper 279, 12 pages. DOI:https://doi.org/10.1145/3290605.3300509

[213] Robert C. Young, Jeffery T. Biggs, Veronika E. Ziegler, and Dolores A. Meyer. 1978. A rating scale for mania: Reliability, validity and sensitivity. *The British Journal of Psychiatry* 133, 5 (1978), 429–435. DOI:https://doi.org/10.1192/bjp.133.5.429

[214] Yakun Yu, Qian Wang, Hao Hu, Shanshan Su, and Zhen Wang. 2018. Multi-atlas based early prediction of post-traumatic stress disorder. In *Proceedings of the 2nd International Symposium on Image Computing and Digital Medicine (ISICDM'2018)*. ACM, 69–72. DOI:https://doi.org/10.1145/3285996.3286012

[215] Liu Yue, Zhang Chunhong, Tian Chujie, Zhao Xiaomeng, Zhang Ruizhi, and Ji Yang. 2018. Application of data mining for young children education using emotion information. In *Proceedings of the 2018 International Conference on Data Science and Information Technology (DSIT'18)*. ACM, 96–104. DOI:https://doi.org/10.1145/3239283.3239321

[216] Sean D. Young, and Renee Garett. 2018. Ethical issues in addressing social media posts about suicidal intentions during an online study among youth: Case study. *JMIR Mental Health* 5, 2 (2018), e33.

[217] Camellia Zakaria, Youngki Lee, and Rajesh Balan. 2019. Passive detection of perceived stress using location-driven sensing technologies at scale (demo). In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys'19)*. ACM, 667–668. DOI:https://doi.org/10.1145/3307334.3328574

[218] Camellia Zakaria, Rajesh Balan, and Youngki Lee. 2019. StressMon: Scalable detection of perceived stress and depression using passive sensing of changes in work routines and group interactions. *Proceedings of the ACM on Human-Computer Interaction 3*, CSCW, Article 37 (November 2019), 29 pages. DOI:https://doi.org/10.1145/3359139

[219] Liang Zhao, Jia Jia, and Ling Feng. 2015. Teenagers' stress detection based on time-sensitive micro-blog comment/response actions. In *Proceedings of the IFIP International Conference on Artificial Intelligence in Theory and Practice*. Springer, 26–38. DOI:https://doi.org/10.1007/978-3-319-25261-2_3

[220] Andrey Zhdanov, Talma Hendler, Leslie Ungerleider, and Nathan Intrator. 2007. Inferring functional brain states using temporal evolution of regularized classifiers. *Intelligent Neuroscience* 2007, Article 52069. DOI:https://doi.org/10.1155/2007/52609

[221] Charles Zheng, Rakesh Achanta, and Yuval Benjamini. 2018. Extrapolating expected accuracies for large multi-class problems. *Journal of Machine Learning Research* 19, 1 (January 2018), 2609–2638.

[222] Dawei Zhou, Jiebo Luo, Vincent Silenzio, Yun Zhou, Jile Hu, Glenn Currier, and Henry Kautz. 2015. Tackling mental health by integrating unobtrusive multimodal sensing. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15)*. AAAI Press, 1401–1408. DOI:https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9546/9334b.

Received September 2019; revised April 2020; accepted May 2020