JMIR AI

Volume 4 (2025) ISSN 2817-1705 Editors-in-Chief: Khaled El Emam, Bradley Malin, PhD

Contents

Review

Original Papers

Research Letter

Review

Identification of Use Cases, Target Groups, and Motivations Around Adopting Smart Speakers for Health Care and Social Care Settings: Scoping Review

Sebastian Merkel[1*], PhD; Sabrina Schorr[1*], MA

Faculty of Social Science, Ruhr University Bochum, Bochum, Germany

[*]all authors contributed equally

Corresponding Author:
Sebastian Merkel, PhD
Faculty of Social Science
Ruhr University Bochum
GD E1/ 155
Universitätsstraße 150
Bochum, 44801
Germany
Phone: 49 0234 32 25411
Email: sebastian.merkel@ruhr-uni-bochum.de

Abstract

Background: Conversational agents (CAs) are finding increasing application in health and social care, not least due to their growing use in the home. Recent developments in artificial intelligence, machine learning, and natural language processing have enabled a variety of new uses for CAs. One type of CA that has received increasing attention recently is smart speakers.

Objective: The aim of our study was to identify the use cases, user groups, and settings of smart speakers in health and social care. We also wanted to identify the key motivations for developers and designers to use this particular type of technology.

Methods: We conducted a scoping review to provide an overview of the literature on smart speakers in health and social care. The literature search was conducted between February 2023 and March 2023 and included 3 databases (PubMed, Scopus, and Sociological Abstracts), supplemented by Google Scholar. Several keywords were used, including technology (eg, voice assistant), product name (eg, Amazon Alexa), and setting (health care or social care). Publications were included if they met the predefined inclusion criteria: (1) published after 2015 and (2) used a smart speaker in a health care or social care setting. Publications were excluded if they met one of the following criteria: (1) did not report on the specific devices used, (2) did not focus specifically on smart speakers, (3) were systematic reviews and other forms of literature-based publications, and (4) were not published in English. Two reviewers collected, reviewed, abstracted, and analyzed the data using qualitative content analysis.

Results: A total of 27 articles were included in the final review. These articles covered a wide range of use cases in different settings, such as private homes, hospitals, long-term care facilities, and outpatient services. The main target group was patients, especially older users, followed by doctors and other medical staff members.

Conclusions: The results show that smart speakers have diverse applications in health and social care, addressing different contexts and audiences. Their affordability and easy-to-use interfaces make them attractive to various stakeholders. It seems likely that, due to technical advances in artificial intelligence and the market power of the companies behind the devices, there will be more use cases for smart speakers in the near future.

(JMIR AI 2025;4:e55673) doi:10.2196/55673

KEYWORDS

conversational agents; smart speaker; health care; social care; digitalization; scoping review; mobile phone

Introduction

Background

In the context of ongoing public debates on artificial intelligence (AI), dialogue systems or conversational agents (CAs) are receiving increasing attention. Their potential applications are being discussed in various fields, including health care [1,2] and social care [3]. CAs have been used in both fields for several years, but recent developments in AI have fueled the scientific discourse [4,5]. The developments in the field of machine learning and natural language processing (NLP), as well as the success of commercially available CAs, such as Amazon's Alexa or Apple's Siri, have been particularly decisive in this regard.

The use of CAs is not limited to a single context; rather, they are used in a variety of settings, including those pertaining to the acquisition of information related to health [6]. CAs using NLP offer a number of features that can be implemented in a variety of health care and social care settings. The field of AI has witnessed considerable progress in recent years, with speech recognition (SR) and NLP advancing significantly. This has enabled the processing of medical terminology in various settings [7]. Although SR in health care has a long tradition dating back to the 1980s, when initial attempts were made to dictate doctor's letters [8], CAs offer multiple additional features. In the context of hands-free interaction, CAs have been used for the purposes of medication reminders [9], symptom management [10], documentation [11], or communication between patients and nurses or doctors, covering multiple medical fields. These include diabetes care [12], monitoring of pregnant women [13], children with special health care needs [11], hearing tests [14], cardiovascular disease [15], and the support of persons with dementia, to name a few [16].

The Rise of Smart Speakers

The term "CA" is not clearly defined, and within the literature, multiple synonyms are used interchangeably. These include "virtual assistants," "AI-driven digital assistants," "voice-based assistants," "voice-controlled intelligent personal assistants," and others. In the study by Laranjo et al [1], the term "CA" is defined as encompassing a range of technologies, including chatbots, embodied CA, which involves a computer-generated character such as an avatar, and smart conversational interfaces, such as Apple's Siri or Amazon's Alexa. In order to characterize CAs, the authors propose that it is necessary to differentiate between the type of technology in question (eg, if the software application is delivered through a mobile device or the telephone), the type of dialogue management (finite-state, frame-based, or agent-based), the actors with control over the dialogue initiative (the user, the system, or a combination of both), the input or output modality (spoken or written, or visual in the case of the output), and whether the system is task-oriented or not [1].

This paper is particularly interested in the use of CAs that are embodied in a physical stationary artifact, which is referred to as a smart speaker. Examples of such devices include Amazon's Echo and Apple's HomePod. Smart speakers are typically confined to a specific location and serve as a platform for a smart conversational interface or AI-driven digital assistant that can be operated through voice input. In the case of the Echo, this is "Alexa", while in the HomePod, it is "Siri". Such assistants are capable of fulfilling a range of tasks, including answering simple questions, switching on lights in conjunction with a smart home system, and playing music. The devices are equipped with one or multiple microphones and software that is capable of analyzing and generating spoken language. In order to operate the devices, the user must utter a designated wake word, such as "Alexa" or "Computer" in the case of Amazon's Echo [17].

The diffusion of smart speakers has been observed to be high in private households in Europe and North America. Amazon launched the first smart speaker in the United States in 2015. As of 2022, approximately 35% of the total US population had used smart speakers [18]. In comparison to the figures from 2019, this represents an increase of 11.1% [19]. A number of studies conducted by market research companies in other countries have reached similar conclusions. For instance, these studies have found that 33% of internet households in the United States, 34% in the United Kingdom [20], and approximately 12%-33% of all households in Germany own at least one smart speaker [21,22].

A recent study by Gaspar and Neus [23] of smart speaker users in the United States, United Kingdom, and Germany shows that Amazon is still the current market leader (United States: 58%; United Kingdom: 71%; and Germany: 68%) followed by Google (United States: 34%; United Kingdom: 22%; and Germany: 25%) and other brands (United States, United Kingdom, and Germany: 7%). It was also found that in all countries, at least 40% (United States: 46%; United Kingdom: 40%; and Germany: 44%) of respondents use smart speakers several times a day. Participants were also asked about the attractiveness of certain application scenarios, including medical diagnosis. Here, participants gave high ratings: United States (19% very attractive and 36% attractive), United Kingdom (12% very attractive and 34% attractive), and Germany (13% very attractive and 35% attractive).

In light of the commercial success of smart speakers and the aforementioned technological advantages in SR and NLP, there has been a growing body of literature on smart speakers in different health care and social care settings [1,24-27]. Commercial devices, such as Amazon's Echo, offer a multitude of features. These devices can be used without any direct contact, are relatively inexpensive and easy to operate, and can be customized and personalized by installing new applications and features [28]. These factors have played a pivotal role in the dissemination of the technology. Finally, the widespread adoption of the technology was driven by the pandemic and the subsequent shift in clinical practices toward greater reliance on digital technologies [29]. Nevertheless, the pervasive use of these devices has also given rise to a multitude of issues and concerns, most notably data collection, storage, and protection [8].

Hence, the devices have attracted increasing attention, with several reviews on CAs in health care settings having been published recently. Each of these reviews has a specific focus:

these include, for instance, design and evaluation challenges [30], effectiveness and usability [31], or chronic conditions [32,33]. To the best of our knowledge, no review has been conducted to date that specifically examines the use of smart speakers within health care and social care settings.

As evidenced by the current state of research, smart speakers are becoming increasingly prevalent in the field of health care and social care. However, there is currently no systematic review available that specifically investigates use cases, settings in which the devices are used, or target groups. To address this gap, our main research question is as follows: What are the scenarios of the use of smart speakers in health care and social care? To address this research question, the main aim of this paper is to present a review of the current research on the use of smart speakers in health care and social care.

Methods

Overview

In order to provide an overview of the existing literature on smart speakers in health care and social care, we conducted a scoping review. The main aim of this approach is to observe, synthesize, and understand current trends [34]. In contrast to a systematic review, which is more suitable for the presentation of a specific clinical question or the presentation of evidence for practice, a scoping review is particularly suitable for identifying features and concepts. Furthermore, it does not aim to provide a synthesizing result for a specific question but rather to provide an overview of a specific topic [34,35]. Thus, the scoping review is a particularly suitable instrument for analyzing the research interest. This encompasses the identification of the nature of the literature, the collation of information on key topics, and the identification of knowledge gaps [35]. Its methodological framework was first published by Arksey and O'Malley [36] and later adapted by Levac, Colquhoun, and O'Brien [37]. Contrary to a systematic review, search terms can be adjusted along the process of a scoping review [36,38]. For the conduction of the present review, the guidelines of Peters et al [39], the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [40] and its extension for Scoping Reviews (PRISMA-ScR) [41] were followed. The results were presented according to the PRISMA checklist (Multimedia Appendix 1).

Search Strategy and Selection Criteria

The literature search was conducted between February 2023 and March 2023. This included a systematic literature search of 3 databases (PubMed, Scopus, and Sociological Abstracts) and a cross-search of the first 20 pages of Google Scholar. This was supplemented by tracing reference lists for further relevant studies. We used the program Citavi 6 for literature management. The review protocol is available on request from the authors. The following keywords were applied in varying combinations and spellings for the systematic search (Table 1):

1. Technology: Here, several terms described above that are found in the literature on CA were used. As the focus of this review is on smart speakers, the search was restricted to this specific type of CA.
2. Product name: As smart speakers were introduced to the market by major American information technology companies, which often use the product names as synonyms for the product, we also included the product or brand names in our search. Globally, Amazon, Google, and Apple are the 3 leading manufacturers; therefore, we included the names of their brands in our search [42].
3. Setting: In order to ensure the most comprehensive search results, we elected to limit our search to the 2 domains of health care and social care without imposing any further restrictions.

Table 1. Keywords used in the literature review.

Technology	Vendor, brand, and product	Setting
Smart speaker	Amazon Alexa	Health care
Voice assistant	Amazon Echo	Social care
Voice-based assistant	Apple HomePod	Care
Voice-controlled assistant	Apple Siri	Nursing
Artificial intelligence–driven digital assistant	Google Home	—[a]
Conversational agent	Google Nest	—
Virtual assistant	—	—

[a]Not applicable.

The terms were linked using Boolean operators. Multiple combinations of the search terms were used using different operators (Multimedia Appendix 2).

To select studies relevant to our research interest, we defined the following inclusion criteria for the full-text screening: (1) publications that were released after 2015, as this was the year in which the first commercial smart speaker was introduced to the market, and (2) the use of a smart speaker in health care and social care settings. No restrictions were placed on the specific setting, including hospitals or long-term care facilities. Furthermore, articles were included in which the devices were not implemented in real settings but were developed for specific settings. Studies were excluded if they met one of the following exclusion criteria: (1) papers that do not report on the specific devices that were used (for instance, in some cases, the authors described the use of a personal assistant without explicitly indicating the specific device on which the assistant was

operational), (2) studies that did not specifically focus on smart speakers (this encompasses the development of voice-operated applications for use on smartphones or tablets), (3) systematic reviews and other forms of literature-based publications, and (4) articles not published in the English language.

Process of Study Selection and Data Extraction

We first screened the titles and abstracts for relevance by both authors. No exclusion criteria were applied to the type of publication during the title and abstract search. Should the title or abstract screening indicate the use of a smart speaker in a health care or social care context, the articles were deemed eligible for full-text screening. For the title and abstract screening, as well as the full-text screening, the same 2 authors reviewed each article independently in order to decide on its inclusion or exclusion. In the event of conflicting decisions regarding inclusion or exclusion, the authors attempted to reach a consensus through discussion. As there was no disagreement, there was no need to involve a third party. The data extraction table contains the following information about each article: (1) authors, (2) year of publication, and (3) country of publication. Furthermore, data were collected on the product and the use case. Furthermore, the following aspects were considered: the settings, the target groups, the motivation for using smart speakers, and the limitations of using such a device. As the primary focus was not on methodological aspects, and due to the heterogeneity of the included literature (some described only technical development while others also included user testing and the often-limited reporting of methods), no such information was collected. The articles included were subjected to qualitative thematic analysis in accordance with the methodology outlined in [43]. Using Kuckartz's [43] approach to qualitative thematic text analysis, researchers identify codes through analysis based on the data gathered. During the process, these codes are then refined. Researchers then identify themes or categories that represent the main findings of the analysis. Identifying themes is a process of examining patterns and similarities between codes and then relating the themes to each other. Consequently, all papers included were read and re-read by both authors, with initial codes being identified. The codes were then compared by the authors, discussed, and grouped into themes. In particular, this included an analysis of the motivation for using the devices and the limitations encountered during the research and development process.

Ethical Considerations

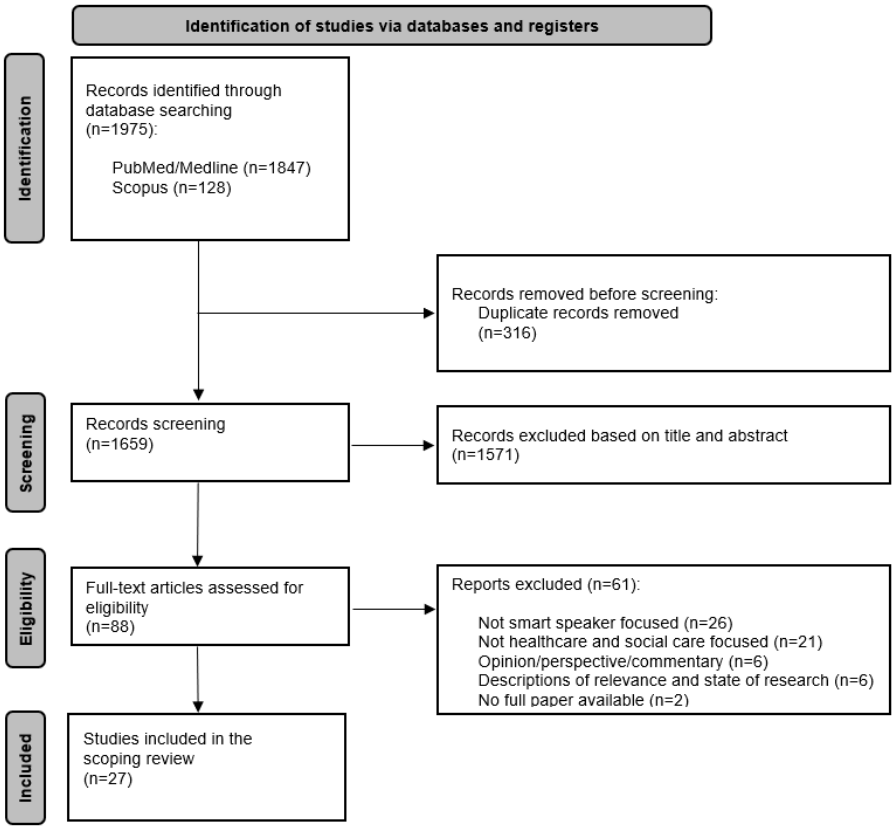
Given the nature of the study, there were no direct interactions with human participants, and thus, no participants to recruit or consent, and no institutional ethical approval was required.

Results

Overview

In total, our search yielded 1975 articles. After removing 316 duplicates, 1659 titles and abstracts were screened by the 2 reviewers. The screening of titles and abstracts resulted in the exclusion of 1571 records, leaving 88 full texts to be assessed for eligibility. Of these, 61 articles were excluded, resulting in a final pool of 27 articles for analysis (Figure 1). The data extraction table for the articles included can be found in Multimedia Appendix 3 [3,9,13-15,44-65].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the search process.



Year and Country of Publication

The majority of articles included in the analysis were published in the United States (n=15 [9, 15, 39, 45, 46, 49, 51, 53-57, 61, 63, 65]), followed by the United Kingdom (n=4 [3,13,43,62]), North Macedonia (n=2 [58,59]), and Australia (n=2 [52,64]). All articles were published between 2018 and 2022, with 2021 being the year with the highest number of publications, with 11 articles.

Technology

There was a clear preference for the devices used: Amazon products were used in 23 of the articles, followed by Google (5). A total of 3 papers used a prototype. It should be noted that some articles used devices from several companies. We found 2 types of articles: Those that use the devices, including the infrastructure (eg, frameworks) provided by the developers, and those that mainly use the hardware (eg, for heart rhythm monitoring; Multimedia Appendix 3 [3,9,13-15,44-65]).

The devices were found to be used In 3 main ways: (1) as standard smart speakers without any further modification, for example, to communicate with patients or to support people living alone (for instance, [44,47]); (2) to develop a skill for a specific use case or multiple use cases (for instance, [48]); and (3) to use the smart speaker and, in some cases, the skill to feed information into another system or as a communication device for other systems (for instance, [15]).

Settings and Target Groups

Given the diverse range of health care and social care settings, we have defined the following categories (Textbox 1). It should be noted that not all articles reported the testing of smart speakers in real health care and social care settings. In some cases, applications were tested in laboratory environments. In the event that this was the case, the intended setting was coded.

Textbox 1. We used the following settings within the domains of health care and special care.

Private homes

- The private living environment includes a person's own home.

Hospitals

- This setting covers acute care hospitals as well as urgent care centers.

Long-term care facilities

- This category includes all settings in which long-term care is provided, for example, nursing homes or rehabilitation centers.

Outpatient services

- This category covers specialized outpatient services, for example, dental or pain management clinics.

Other

- In case the device was tested in a setting not matching the definition of the ones listed above, we categorized it as "other." For instance, this could be in a car.

Furthermore, 4 target groups were identified. It should be noted that an article can have several target groups, including (1) patients, (2) medical staff members such as physicians, (3) nurses and professional caregivers, and (4) informal caregivers who provide unpaid help to a friend or family member. Moreover, category (5), "other," was defined for all target groups not matching any of the aforementioned. It should be noted that multiple target groups were covered in one article. Only those who directly interact with the device were included. For instance, Domínguez et al [50] developed a system to support assisted reproduction treatment. Although physicians are involved, only the patients interact with a smart speaker and hence were included.

The most prevalent setting mentioned in the studies included was home care (n=20), followed by hospitals (n=6). Outpatient care (n=3) was less frequently observed (Multimedia Appendix 3 [3,9,13-15,44-65]). In one instance, the setting was not specified [14]. However, it is best classified under home care.

Among the target groups, patients are the most frequent users mentioned in 23 of the articles (Multimedia Appendix 3 [3,9,13-15,44-65]). Older adults, in particular, were often seen as a promising target group, and we found that 11 of the included publications focus on this target group [66] (Multimedia Appendix 3 [3,9,13-15,44-65]). While some articles included descriptions of the development and testing of skills specifically designed for older adults [51,52], others explored the general acceptance and potential of the technology for older adults. For instance, Lee et al [51] developed multiple skills aimed at older persons, including a reminder to take medication, a diet tracking system, and a skill alerting caregivers in case of a fall. Nallam et al [49] simulated a CA to answer health-related questions asked by older persons. O'Brien et al [47] used off-the-shelf devices without any form of modification to investigate the effects on home-bound older adults with social isolation. The participants used the devices for a variety of purposes, including monitoring their health and well-being, as well as for emergency communication. Some authors report that older adults constitute the largest group of first adopters of smart speakers. In addition, smart speakers allow easy contact with caregivers [12] or low-threshold access to health information [13]. Older adults as potential users of CA have been the focus before [39,67,68].

The second most frequent target group was physicians (n=11), followed by other health professionals (eg, nurses; n=9) and informal caregivers (n=1; Multimedia Appendix 3 [3,9,13-15,44-65]). These results demonstrate that the majority of articles focus on supporting nonresidential care.

Table 2 provides an overview of all settings and target groups. It is important to note that a single paper can include multiple settings and target groups.

While previous statements covered the number of papers included in the review, Table 2 combines user groups and settings across the studies covered. It shows that patients are the most common target group, while home care is the most common setting.

Table 2. Settings and user groups.

	Patients	Physicians	Older adults	Nurses and so on	Informal caregivers	Other	Total
Home care	19	5	11	7	1	1	44
Hospitals	4	5	0	2	0	0	11
Outpatient care	2	2	1	1	0	0	6
Total	25	12	12	10	1	1	

Use Cases

We found several use cases covering, among others, hearing tests [14], cardiovascular diseases [15,46], pregnancy companion [13], cancer management [eg, 58,59], or medication reminders [69]. It must be noted that several articles reported that smart speakers were used in multiple use cases. For example, Wright [70] describes that a local authority was involved in developing applications, including “a Skill that prompted users to take their medicine; a Skill that helped to record and manage care tasks; a Skill to facilitate communication with caregivers by recording messages; and a Skill to connect users to a trusted LA directory of services” [44]. Jadczyk et al [71], who developed a voice-enabled automated platform for the collection of medical data from patients with cardiovascular disease, describe 5 use cases within their study: (1) education, (2) process optimization, (3) patient support, and (4) data collection, and (5) medical device grade solutions (eg, diagnose and treatment). The devices were used to open patient files and images, initiate conference calls, or record images and videos [4].

While most of the identified use cases were found in the domain of health care, social care played a subordinate role. Still, we found several articles reporting on the use of smart speakers in this domain. Within this field, elderly care was the most relevant area. For instance, O’Brien et al [47] use a smart speaker to reduce loneliness and social isolation among older adults living at home. Palumbo et al [72] developed personalized coaching for older individuals to increase their well-being by aiming at the areas of physical activity, nutrition, cognition, and social relationships. In the domain of social care, older adults living at home or care home residents were the main user group (eg, [3]).

Motivation for Use

The reasons for using smart speakers in health care are framed with various arguments. Besides their low acquisition costs [51], this also includes aspects applying to digital technologies in health care and social care in general, such as the possibility to deliver care remotely without restrictions in time and space (eg, Sadavarte et al [13]). Another motivation is the fact that smart speakers are already widely accepted as a consumer technology [45,52]. Hence, users already know how to operate the devices and are also familiar with their limitations. Other aspects cover potentially increased productivity across the use cases that we identified. For instance, Bhatt et al [45] used a voice-based assistant to access and update an electronic health record. They see advantages in terms of efficiency (less time spent on data input) and accuracy, as speech-to-text might result in fewer errors. Ultimately, this might also benefit patients as waiting time is reduced [45]. Jadczyk et al [71] highlighted the main potential in the possibility of automating traditional telehealth services: “Voice chatbots can support routine care through automatic at-home monitoring, triaging, screening, providing medical recommendations and guidelines, and improving operational workflow” [15,71].

Another advantage is the user interface, which is easy to navigate [11]. Cheng et al [55] argue that the main advantage of the technology is that it: “eliminate[s] the struggles that are associated with strictly tactile screens.” (2018); or that human-like verbal communication that feels more natural and intuitive and particularly that the devices can be used hands-free [55]. Jansons et al [52] drive on the research of Foehr and Germelmann [73] and argue that the devices “may enhance adherence to remotely-delivered exercise interventions […], because the human-like attributes associated with these technologies may elicit a sense of familiarity, social presence, and human engagement” [52]. Moreover, the authors see this as an advantage for older users [53] who support this viewpoint and argue that “digital non-natives” might be especially benefitting from this technology. For instance, Kim [4] tested the experiences of older adults who used the devices for the first time and found that due to the simple interaction, health-related questions were a typical use case.

The form of smart speakers and their design were mentioned in some publications. Gouda et al [74] saw the fact that smart speakers are “non-invasive” technology as a main advantage. As the devices can be placed nearly anywhere in the room and can be operated without the need to see them, it allows for new ways of interaction. Luo et al [56] also see a benefit in the fact that the immobility of the devices is as helpful as this helps, in contrast to mobile phones, in establishing habits and routines.

Wright [44] describes the use of smart speakers in trials run by local authorities in England. Drawing on interviews with managers from 8 English local authorities, benefits are seen in the low-cost supplement or alternative to telecare. Or, as one of his interview partners put it: "have the advantages of being sophisticated and powerful, relatively cheap, already widely used and familiar, designed with a degree of accessibility and intuitive use in mind, and a growing level of interoperability with other networked digital devices aided by an open development framework" [44]. One of the results of the study is that local authorities chose Amazon's Echo because of "councils facing depleted funds, a lack of expert guidance on care technologies, and an increasingly complex and fragmented care technology marketplace" [44].

Limitations of Smart Speakers

In addition, various limitations of the technology were addressed in the included articles. Here, most technical limitations were named (1) insufficient hearing comprehension [57], speech recognition [51], or emotion recognition [54]; (2) that there is no interruption of the recording during slow speeches allowed [14]; (3) difficult functioning in the natural living environments due to interfering noises [3]; (4) that the correctness of the answer is not always accurate [51]; and (5) that the devices allow longer conversations [49]. Internet access must also be provided [48,75]. Besides these technical aspects, there were also social aspects mentioned. This covered the (lack of) user acceptance, particularly among older users and professional caregivers [45,76], but also their lack of basic digital skills [75]. These supposedly low digital skills might lead to challenges in interacting with the devices. Users might forget the wake word, there may be timing issues when communicating with the devices, or they might have difficulties in setting up the devices [47,53]. Another issue that was mentioned regularly was data protection. Here, the misuse of sensitive data is particularly pointed out. For example, if security measures are inadequate, it would be possible to manipulate the medication and thus actively harm the patient [12]. Cheng et al [55] also argue for multimodal solutions as people might feel uncomfortable talking to devices in front of other people.

Discussion

Principal Findings

Our aim was to identify use cases and scenarios in which smart speakers can be used within health care and social care. The results show that smart speakers are used in various contexts and for multiple reasons. The main features used are NLP and hands-free interaction. Moreover, the fact that the technology is widely used in private homes and hence many persons are used to interact with the devices are important aspects. In addition to offering relatively inexpensive hardware, smart speakers and the companies behind them provide software frameworks and infrastructure, such as Amazon's skill, which assists developers in the design and marketing of their products.

It is important to note that there is no clear definition of smart speakers. One challenge of this study was the varying definitions of the technology, with the term often being used interchangeably with personal assistants such as Siri or Cortana. These assistants play an important role in the use of smart speakers, which arguably only serve as a shell equipped with microphones and loudspeakers for them. However, we argue that smart speakers should be considered a distinct technology. Based on this review, we understand smart speakers as a type of CA bound to a fixed location. Within the field of health care and social care, the technology can be used in various settings and use cases such as communication, documentation, or diagnosis and therapy of diseases hands-free. Smart speakers are equipped with microphones and loudspeakers and connected to the internet. They usually come with an integrated digital assistant, but even without such an assistant, they offer multiple features that can be used across various settings. Smart speakers can be customized using either skills or apps that can be installed on the devices.

The results show that all publications were published between 2018 and 2021. Furthermore, the majority were published in the United States. The following explanations can be given for these 2 results. Alexa was the first voice assistant that was compliant with the Health Insurance Portability and Accountability Act (HIPAA), allowing it to be the access example of clinical records. In England, the National Health Service contracted with Amazon to enable Alexa in 2019 to answer health-related questions, raising questions about privacy and how health care data would be used [44,45]. The HIPAA compliance and the fact that the National Health Service contracted with Amazon explains why most studies have been carried out in the United States and the United Kingdom. Arguably, European countries are not as present due to more strict data protection regulations. Moreover, the use of smart speakers is significantly higher in the United States than in other countries, which in turn could also be related to data protection regulations [77]. Interestingly, Asian countries have, with few exceptions, also not been represented in the included articles. This seems counterintuitive as, in terms of market sales, smart speaker technology by Asian technology companies is more and more successful [42].

It also became clear that the devices were clearly dominant in the publications. This should be criticized from a scientific point of view. We were able to identify the following explanations for this result.

Since Amazon entered the market in 2015 and continuously updates its product line, off-the-shelf devices have recently increased in terms of market penetration, making them more popular for research and development. That Amazon's Echo was used in the vast majority of articles included comes as no surprise, and Amazon's market dominance is based on several factors. First, the company was the first to release a smart speaker to consumers. Second, Amazon's voice assistant, Alexa, has been embedded in a broad range of devices, including wall clocks, by third-party manufacturers. Third, Amazon sells products of the Echo family at comparably low prices, starting at around US $20. Fourth, Amazon offers an infrastructure through its Skill Store and several frameworks for developers. Fifth, in the United States, the Echo is HIPAA-compliant.

The dominance of Amazon's smart speaker in the included papers poses several risks depending on the use case, some of

which are discussed in the papers themselves. In terms of the devices themselves in their off-the-shelf version, the interaction is limited. For example, Nallam et al [49] used a smart speaker prototype as they argue that developed solutions often do not support conversational interactions and explore scenarios that are not yet supported.

The articles included in this publication address a diverse range of use cases across various settings, thereby demonstrating the versatility of smart speakers and the technology of NLP and AI incorporated in them. This technology can be used in a multitude of contexts within the domains of health care and social care. Overall, 2 general use cases can be distinguished: (1) supporting patients and their relatives in their private living environments and (2) supporting professional health care workers in clinical settings. As the devices were originally developed for private home environments and primarily for entertainment and e-commerce applications, it is unsurprising that this setting was the dominant one across the papers included in this review. This could be seen as an indicator of the restructuring of health care services, with an increased focus on the private living environment. Several clinical use cases supported by smart speakers could be automated and not be restricted to clinical settings (eg, [14,48]). Only in a few cases does the paper focus on clinical use cases and professional personnel (eg, [4,45,71]).

That patients, and particularly older adults, were the main target group supports this conclusion. Moreover, this also underlines that the role of patients and practices of health and care change against the background of digitalization and the use of AI [78]. While some of the use cases identified were exclusively designed for clinical settings, the majority can, in theory, be implemented in multiple settings. This could support patient empowerment, as smart speakers can be used to support the household as a central place of health care. An argument supporting the fit of the devices for older adults is that smart speakers do not require "reasonable levels of vision and manual dexterity" [79,80].

A key rationale for using the devices is not only their competitive pricing but also the potential to reduce expenditure by enhancing the efficiency of staff members and care processes, for instance, through enhanced documentation or facilitating straightforward communication with patients, colleagues, or clients. Although the majority of the papers reviewed argue that smart speakers could provide such benefits, these potential benefits depend on several circumstances. The first is whether the devices can be installed as they are or whether new skills or, more complexly, additional hardware or modifications are required. This depends on the use case and also the target group. Although many people are used to interacting with the devices, older adults might not have any experience and could need training.

The majority of the papers in our sample can be classified as exploratory in nature. The research designs used are predominantly qualitative, with sample sizes that are relatively small and no long-term studies conducted in real-world scenarios. This underscores the fact that the technology itself is still relatively new, particularly within the context of health care and social care. In addition, researchers and developers are still exploring the technology's potential applications in health care and social care, which may have become more apparent in the context of the pandemic. Both sectors are currently experiencing financial strain due to rising expenditure and a shortage of qualified personnel [81]. New technologies are frequently viewed as a potential solution to these challenges [70].

Smart speakers and digital voice assistants like Alexa are quite limited in terms of their initial dialogue management, which can be seen as an important motivator to using the systems as they are easier to develop and control. This finding is in line with a systematic review of CA in health care carried out by Laranjo et al [1]. The authors could identify 17 articles using 14 different CA. Most papers covered by the review evaluated task-oriented CA that aims at supporting patients and clinicians. Systems allowing the management of complex dialogues were only identified in 1 case. Even though conversational systems have proven to be beneficial for health-related purposes, most assistants allow only constrained user input (eg, multiple-choice answers) [1,82]. Clark et al [83] argue that users interact in "clearly delineated task-based conversations" and "fall short of reflexive and adaptive interactivity." According to the authors, the term conversation is "a poor description of the current interaction experience" with an AI using common smart speakers [83]. Hence, they suggest testing "human-agent interaction as a new genre of conversation, with its own rules, norms and expectations" [83]. The devices have only a limited capability to actually be able to engage in a conversational dialogue. Conversations are task-oriented instead of offering interactions initiated by the user and not by the device. While this might be true, it seems to be only a matter of time before future updates might be used to allow more natural dialogues, as is already the case with generative AI such as ChatGPT.

The analysis showed that change in existing practices and routines is an important aspect. Drawing on Sezgin et al [84], Capasso and Umbrello [85] argue that the novelty of CAs is that they act as "intermediaries between the health care system as a whole and the public," changing practices in health care and social care. Here, several studies follow the normative aim to implement innovative technologies in order to improve processes and outcomes. The use of smart speakers—or CAs in general—follows a technology-driven approach. Already existing technologies are transferred to the domains of health and social care. Due to the exploratory design of most studies, the emphasis is put on the technology and not on the context, like organizational or social factors. The logic of a "fitting" technology seems to be a main driver of many studies, neglecting the analysis of potentially changing social practices.

The dominance of Amazon in our sample has to be seen from a critical perspective. The company itself began offering the service Alexa Together and was able to emulate existing approaches and leverage its financial and market clout to challenge competitors. Moreover, developers depend on the technology, that is, the hardware and also the software frameworks of one company. As a consequence, the dominant position of Amazon might increase due to research using the company's products. If only one product from a particular company is examined, the capabilities of other products are not

taken into account, as they may perform better, for example, and might be used to copy promising applications.

Limitations

This paper has several limitations. First, the number of databases searched. To address this limitation, a cross-search was performed in Google Scholar to rule out the possibility that important articles were not found. In addition, to broaden the search strategy, other forms of literature, such as trial reports, could be included in future studies. For instance, a few trials using smart speakers are registered on clinicaltrials.gov. However, we decided not to include these as they did not provide all the information we wanted to obtain (eg, motivations for using the devices). Second, we restricted our search to the English language only. Few papers were found from the Asian region, probably due to the language limitation of the search. This limitation was mitigated by using brand names as search terms focusing on the brands with the highest market share. However, as recent market research shows, there is a shift toward products developed in Asian countries, and future studies should include a wider range of brands and products. Another limitation is that we only looked at smart speakers, which excludes other voice assistants that use essentially the same technology (such as digital assistants on smartphones and tablets). We deliberately excluded these as this review focused specifically on smart speakers as a form of CA, and we argue that the technology of smart speakers needs to be seen as a technology in its own right.

Conclusion

In this paper, a scoping review was conducted on the use of smart speakers in health care and social care settings. The analysis showed that—due to the widespread use of devices like Amazon's Echo—smart speaker technology has been tested and implemented in various settings and use cases in the health and social care sectors. The main setting was the private home environment, and the main user group was patients. There are, however, also approaches to making use of the technology in other settings, such as hospitals. It seems likely that due to technical progress in the field of AI and the market power of the companies behind the devices, there will be more use cases of smart speakers in the (near) future.

Acknowledgments

This study was supported by the Federal Ministry of Education and Research (grant number 16SV8791).

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist.
[DOCX File , 22 KB - ai_v4i1e55673_app1.docx]

Multimedia Appendix 2
Database search details.
[DOCX File , 18 KB - ai_v4i1e55673_app2.docx]

Multimedia Appendix 3
Data extraction table.
[DOCX File , 38 KB - ai_v4i1e55673_app3.docx]

References

1. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
2. Provoost S, Lau HM, Ruwaard J, Riper H. Embodied conversational agents in clinical psychology: a scoping review. J Med Internet Res 2017;19(5):e151 [FREE Full text] [doi: 10.2196/jmir.6553] [Medline: 28487267]
3. Edwards KJ, Jones RB, Shenton D, Page T, Maramba I, Warren A, et al. The use of smart speakers in care home residents: implementation study. J Med Internet Res 2021;23(12):e26767 [FREE Full text] [doi: 10.2196/26767] [Medline: 34932010]
4. Kim S. Exploring how older adults use a smart speaker-based voice assistant in their first interactions: qualitative study. JMIR mHealth uHealth 2021;9(1):e20427 [FREE Full text] [doi: 10.2196/20427] [Medline: 33439130]

5. Tudor Car L, Dhinagaran DA, Kyaw BM, Kowatsch T, Joty S, Theng Y, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020;22(8):e17158 [FREE Full text] [doi: 10.2196/17158] [Medline: 32763886]
6. Ermolina A, Tiberius V. Voice-controlled intelligent personal assistants in health care: international Delphi study. J Med Internet Res 2021;23(4):e25312 [FREE Full text] [doi: 10.2196/25312] [Medline: 33835032]
7. Chung AE, Griffin AC, Selezneva D, Gotz D. Health and fitness apps for hands-free voice-activated assistants: content analysis. JMIR mHealth uHealth 2018;6(9):e174 [FREE Full text] [doi: 10.2196/mhealth.9705] [Medline: 30249581]
8. Kumah-Crystal YA, Pirtle CJ, Whyte HM, Goode ES, Anders SH, Lehmann CU. Electronic health record interactions through voice: a review. Appl Clin Inform 2018;9(3):541-552 [FREE Full text] [doi: 10.1055/s-0038-1666844] [Medline: 30040113]
9. Corbett CF, Combs EM, Chandarana PS, Stringfellow I, Worthy K, Nguyen T, et al. Medication adherence reminder system for virtual home assistants: mixed methods evaluation study. JMIR Form Res 2021;5(7):e27327 [FREE Full text] [doi: 10.2196/27327] [Medline: 34255669]
10. Vuppalapati JS, Kedari S, Ilapakurti A, Kedari S, Gudivada M, Vuppalapati C. The role of voice service technologies in creating the next generation outpatient data driven electronic health record (EHR). : IEEE; 2017 Presented at: 2017 Intelligent Systems Conference (IntelliSys); September 7-8, 2017; London, United Kingdom. [doi: 10.1109/intellisys.2017.8324289]
11. Sezgin E, Noritz G, Elek A, Conkol K, Rust S, Bailey M, et al. Capturing at-home health and care information for children with medical complexity using voice interactive technologies: multi-stakeholder viewpoint. J Med Internet Res 2020;22(2):e14202 [FREE Full text] [doi: 10.2196/14202] [Medline: 32053114]
12. Basatneh R, Najafi B, Armstrong DG. Health sensors, smart home devices, and the internet of medical things: an opportunity for dramatic improvement in care for the lower extremity complications of diabetes. J Diabetes Sci Technol 2018;12(3):577-586 [FREE Full text] [doi: 10.1177/1932296818768618] [Medline: 29635931]
13. Sadavarte SS, Bodanese E. Pregnancy companion chatbot using Alexa and Amazon Web Services. : IEEE; 2019 Presented at: 2019 IEEE Pune Section International Conference (PuneCon); December 18-20, 2019; Pune, India. [doi: 10.1109/punecon46936.2019.9105762]
14. Ooster J, Moreta PNP, Bach JH, Holube I, Meyer BT. 'Computer, Test My Hearing': accurate speech audiometry with smart speakers. 2019 Presented at: Interspeech 2019; 2019 September 15-19; Graz, Austria. [doi: 10.21437/interspeech.2019-2118]
15. Jadczyk T, Kiwic O, Khandwalla RM, Grabowski K, Rudawski S, Magaczewski P, et al. Feasibility of a voice-enabled automated platform for medical data collection: cardioCube. Int J Med Inform 2019;129:388-393. [doi: 10.1016/j.ijmedinf.2019.07.001] [Medline: 31445282]
16. Rampioni M, Stara V, Felici E, Rossi L, Paolini S. Embodied conversational agents for patients with dementia: thematic literature analysis. JMIR mHealth uHealth 2021;9(7):e25381 [FREE Full text] [doi: 10.2196/25381] [Medline: 34269686]
17. Waldhör K. Smarte objekte – wie smart speaker und smarthome die medizinische und pflegerische versorgung zu hause unterstützen werden [Book in German]. In: Digitale Transformation von Dienstleistungen im Gesundheitswesen VI. Wiesbaden: Springer Fachmedien Wiesbaden; 2019:389-406.
18. Smart speakers statistics: report 2022. Speakergy. 2022. URL: https://speakergy.com/smart-speakers-statistics/#:~:text=The%20United%20States%20Smart%20Speaker,a%206%25%20increase%20from%202020 [accessed 2022-09-30]
19. Petrock V. Voice assistant and smart speaker users 2020: more time at home means more time to talk. 2020. URL: https://www.emarketer.com/content/voice-assistant-and-smart-speaker-users-2020 [accessed 2021-12-02]
20. INSIGHTS 2020: device usage 2020. AudienceProject. 2020. URL: https://www.audienceproject.com/wp-content/uploads/audienceproject_study_device_usage_2020.pdf [accessed 2021-12-02]
21. Initiative D21 e. V. D21-Digital-Index 2021/2022 [Website in German]. Jährliches Lagebild zur Digitalen Gesellschaft. 2022. URL: https://initiatived21.de/publikationen/d21-digital-index/2021-2022 [accessed 2024-12-10]
22. Welcome to 'The Age of Voice 3.0': OMD Germany. OMD. 2021. URL: https://www.omd.com/news/welcome-to-the-age-of-voice-3-0/ [accessed 2023-02-18]
23. Gaspar C, Neus A. Smart-speaker-report 2023: erfahrungen, bewertungen und wünsche der nutzer in Deutschland, UK und Den USA [Article in German]. Nürnberg Institut für Marktentscheidungen e.V. 2023. URL: https://www.nim.org/publikationen/detail/smart-speaker-report-2023 [accessed 2024-12-10]
24. Baertsch MA, Decker S, Probst L, Joneleit S, Salwender H, Frommann F, et al. Convenient access to expert-reviewed health information via an alexa voice assistant skill for patients with multiple myeloma: development study. JMIR Cancer 2022;8(2):e35500 [FREE Full text] [doi: 10.2196/35500] [Medline: 35679096]
25. Beaman J, Lawson L, Keener A, Mathews ML. Within clinic reliability and usability of a voice-based Amazon Alexa administration of the patient health questionnaire 9 (PHQ 9). J Med Syst 2022;46(6):38 [FREE Full text] [doi: 10.1007/s10916-022-01816-0] [Medline: 35536347]
26. Brewer RN. 'If Alexa knew the state I was in, it would cry': older adults' perspectives of voice assistants for health. 2022 Presented at: CHI Conference on Human Factors in Computing Systems Extended Abstracts; April 29, 2022; New Orleans, LA, USA p. 1-8. [doi: 10.1145/3491101.3519642]
27. Sunshine J. Smart speakers: the next frontier in mHealth. JMIR mHealth uHealth 2022;10(2):e28686. [doi: 10.2196/28686] [Medline: 35188467]

28. Kocaballi AB, Berkovsky S, Quiroz JC, Laranjo L, Tong HL, Rezazadegan D, et al. The personalization of conversational agents in health care: systematic review. J Med Internet Res 2019;21(11):e15360 [FREE Full text] [doi: 10.2196/15360] [Medline: 31697237]
29. Kraus S, Schiavone F, Pluzhnikova A, Invernizzi AC. Digital transformation in healthcare: analyzing the current state-of-research. J Bus Res 2021;123:557-567. [doi: 10.1016/j.jbusres.2020.10.030]
30. Kocaballi AB, Sezgin E, Clark L, Carroll JM, Huang Y, Huh-Yoo J, et al. Design and evaluation challenges of conversational agents in health care and well-being: selective review study. J Med Internet Res 2022;24(11):e38525 [FREE Full text] [doi: 10.2196/38525] [Medline: 36378515]
31. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020;22(10):e20346 [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]
32. Bin Sawad A, Narayan B, Alnefaie A, Maqbool A, Mckie I, Smith J, et al. A systematic review on healthcare artificial intelligent conversational agents for chronic conditions. Sensors (Basel) 2022;22(7):2625 [FREE Full text] [doi: 10.3390/s22072625] [Medline: 35408238]
33. Schachner T, Keller R, V Wangenheim F. Artificial intelligence-based conversational agents for chronic conditions: systematic literature review. J Med Internet Res 2020;22(9):e20701 [FREE Full text] [doi: 10.2196/20701] [Medline: 32924957]
34. Jahan N, Naveed S, Zeshan M, Tahir MA. How to conduct a systematic review: a narrative literature review. Cureus 2016;8(11):e864 [FREE Full text] [doi: 10.7759/cureus.864] [Medline: 27924252]
35. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
36. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
37. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
38. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
39. O'Brien K, Liggett A, Ramirez-Zohfeld V, Sunkara P, Lindquist LA. Voice-controlled intelligent personal assistants to support aging in place. J Am Geriatr Soc 2020;68(1):176-179. [doi: 10.1111/jgs.16217] [Medline: 31617581]
40. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
41. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
42. Strategy analytics: global smart speaker shipments declined 5% in 1Q22 amid disruption from war and a resurgent COVID virus. Businesswire. 2022. URL: https://www.businesswire.com/news/home/20220606005136/en/Strategy-Analytics-Global-Smart-Speaker-Shipments-Declined-5-in-1Q22-Amid-Disruption-from-War-and-a-Resurgent-COVID-Virus [accessed 2022-09-30]
43. Kuckartz U. Qualitative Text Analysis: A Systematic Approach. In: Compendium for Early Career Researchers in Mathematics Education. Cham: Springer Nature; 2019:181-197.
44. Wright J. The Alexafication of adult social care: virtual assistants and the changing role of local government in England. Int J Environ Res Public Health 2021;18(2):812 [FREE Full text] [doi: 10.3390/ijerph18020812] [Medline: 33477872]
45. Bhatt V, Li J, Maharjan B. DocPal: a voice-based EHR assistant for health practitioners. : IEEE; 2021 Presented at: 2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM); 2021 March 01-02; Shenzhen, China. [doi: 10.1109/healthcom49281.2021.9399013]
46. Wang A, Nguyen D, Sridhar AR, Gollakota S. Using smart speakers to contactlessly monitor heart rhythms. Commun Biol 2021;4(1):319 [FREE Full text] [doi: 10.1038/s42003-021-01824-9] [Medline: 33750897]
47. O'Brien K, Light SW, Bradley S, Lindquist L. Optimizing voice-controlled intelligent personal assistants for use by home-bound older adults. J Am Geriatr Soc 2022;70(5):1504-1509 [FREE Full text] [doi: 10.1111/jgs.17625] [Medline: 35029296]
48. Sharma A, Oulousian E, Ni J, Lopes R, Cheng MP, Label J, et al. Voice-based screening for SARS-CoV-2 exposure in cardiovascular clinics. Eur Heart J Digit Health 2021;2(3):521-527 [FREE Full text] [doi: 10.1093/ehjdh/ztab055] [Medline: 36713601]
49. Nallam P, Bhandari S, Sanders J, Martin-Hammond A. A question of access: exploring the perceived benefits and barriers of intelligent voice assistants for improving access to consumer health resources among low-income older adults. Gerontol Geriatr Med 2020;6:2333721420985975 [FREE Full text] [doi: 10.1177/2333721420985975] [Medline: 33457459]

50. Domínguez D, Morales L, Sánchez N. IoMT-Driven eHealth: a technological innovation proposal based on smart speakers. In: Rojas I, Valenzuela O, Rojas F, editors. Bioinformatics and Biomedical Engineering. Cham: Springer International Publishing; 2020:378-386.
51. Lee E, Vesonder G, Wendel E. Eldercare robotics - Alexa. : IEEE; 2020 Presented at: 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON); October 28-31, 2020; New York, NY, USA p. 820-825. [doi: 10.1109/uemcon51285.2020.9298147]
52. Jansons P, Fyfe J, Via JD, Daly RM, Gvozdenko E, Scott D. Barriers and enablers for older adults participating in a home-based pragmatic exercise program delivered and monitored by Amazon Alexa: a qualitative study. BMC Geriatr 2022;22(1):248 [FREE Full text] [doi: 10.1186/s12877-022-02963-2] [Medline: 35337284]
53. Qiu L, Kanski B, Doerksen S, Winkels RM, Schmitz K, Abdullah S. Nurse AMIE: using smart speakers to provide supportive care intervention for women with metastatic breast cancer. : ACM; 2021 Presented at: CHI '21: CHI Conference on Human Factors in Computing Systems; 2021 May 8-13; Yokohama Japan. [doi: 10.1145/3411763.3451827]
54. Thomas G. Patient and clinician-centric healthcare enhancement through speech recognition: a research proposal. 2019 Presented at: 7th Annual International Conference on Architecture and Civil Engineering (ACE 2019) GSTF 2019; May 27-28, 2019; Singapore URL: https://dl4.globalstf.org/products-page/books/patient-and-clinician-centric-healthcare-enhancement-through-speech-recognition/ [doi: 10.5176/2301-394X_ACE19.581]
55. Cheng A, Raghavaraju V, Kanugo J, Handrianto YP, Shang Y. Development and evaluation of a healthy coping voice interface application using the Google home for elderly patients with type 2 diabetes. : IEEE; 2018 Presented at: 2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC); January 12-15, 2018; Las Vegas, NV, USA p. 1-5. [doi: 10.1109/ccnc.2018.8319283]
56. Luo Y, Lee B, Choe E. TandemTrack: shaping consistent exercise experience by complementing a mobile app with a smart speaker. : ACM; 2020 Presented at: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020 April 25-30; Honolulu HI USA p. 1-13. [doi: 10.1145/3313831.3376616]
57. Arem H, Scott R, Greenberg D, Kaltman R, Lieberman D, Lewin D. Assessing breast cancer survivors' perceptions of using voice-activated technology to address insomnia: feasibility study featuring focus groups and in-depth interviews. JMIR Cancer 2020;6(1):e15859 [FREE Full text] [doi: 10.2196/15859] [Medline: 32348274]
58. Dojchinovski D, Ilievski A, Gusev M. Interactive home healthcare system with integrated voice assistant. 2019 Presented at: 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO); May 20-24, 2019; Opatija, Croatia URL: https://ieeexplore.ieee.org/document/8756983 [doi: 10.23919/MIPRO.2019.8756983]
59. Ilievski A, Dojchinovski D, Gusev M. Interactive voice assisted home healthcare systems. New York, NY: Association for Computing Machinery; 2019 Presented at: BCI'19: 9th Balkan Conference in Informatics; September 26-28, 2019; Sofia, Bulgaria. [doi: 10.1145/3351556.3351572]
60. Yoo TK, Oh E, Kim H, Ryu IH, Lee IS, Kim JS, et al. Deep learning-based smart speaker to confirm surgical sites for cataract surgeries: a pilot study. PLoS One 2020;15(4):e0231322 [FREE Full text] [doi: 10.1371/journal.pone.0231322] [Medline: 32271836]
61. Ismail HO, Moses AR, Tadrus M, Mohamed EA, Jones LS. Feasibility of use of a smart speaker to administer Snellen visual acuity examinations in a clinical setting. JAMA Netw Open 2020 Aug 03;3(8):e2013908 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.13908] [Medline: 32822489]
62. Chambers R, Beaney P. The potential of placing a digital assistant in patients' homes. Br J Gen Pract 2020 Jan;70(690):8-9 [FREE Full text] [doi: 10.3399/bjgp20X707273] [Medline: 31879289]
63. Kim JH, Um R, Liu J, Patel J, Curry E, Aghabaglou F, et al. Development of a smart hospital assistant: integrating artificial intelligence and a voice-user interface for improved surgical outcomes. Proc SPIE Int Soc Opt Eng 2021 Feb;11601:116010U [FREE Full text] [doi: 10.1117/12.2580995] [Medline: 35341075]
64. Jansons P, Dalla Via J, Daly RM, Fyfe JJ, Gvozdenko E, Scott D. Delivery of home-based exercise interventions in older adults facilitated by Amazon Alexa: a 12-week feasibility trial. J Nutr Health Aging 2022;26(1):96-102 [FREE Full text] [doi: 10.1007/s12603-021-1717-0] [Medline: 35067710]
65. Apergi LA, Bjarnadottir MV, Baras JS, Golden BL, Anderson KM, Chou J, et al. Voice interface technology adoption by patients with heart failure: pilot comparison study. JMIR mHealth uHealth 2021 Apr 01;9(4):e24646 [FREE Full text] [doi: 10.2196/24646] [Medline: 33792556]
66. Martin-Hammond A, Vemireddy S, Rao K. Exploring older adults' beliefs about the use of intelligent assistants for consumer health information management: a participatory design study. JMIR Aging 2019;2(2):e15381 [FREE Full text] [doi: 10.2196/15381] [Medline: 31825322]
67. Bickmore TW, Caruso L, Clough-Gorr K. Acceptance and usability of a relational agent interface by urban older adults. 2005 Presented at: Extended Abstracts Proceedings of the 2005 Conference on Human Factors in Computing Systems, CHI 2005; 2005 April 2-7; Portland, Oregon, USA p. 1212-1215. [doi: 10.1145/1056808.1056879]
68. Vardoulakis LP, Ring L, Barry B, Sidner CL, Bickmore T. Designing Relational Agents as Long Term Social Companions for Older Adults. In: Hutchison D, Kanade T, Kittler J, editors. Intelligent Virtual Agents. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012:289-302.

69. Jesús-Azabal M, Medina-Rodríguez J, Durán-García J, García-Pérez D. Remembranza Pills: Using Alexa to Remind the Daily Medicine Doses to Elderly. In: García-Alonso J, Fonseca C, editors. Gerontechnology. Cham: Springer International Publishing; 2020:151-159.
70. Henwood F, Marent B. Understanding digital health: productive tensions at the intersection of sociology of health and science and technology studies. Sociol Health Illn 2019;41 Suppl 1:1-15. [doi: 10.1111/1467-9566.12898] [Medline: 31599984]
71. Jadczyk T, Wojakowski W, Tendera M, Henry TD, Egnaczyk G, Shreenivas S. Artificial intelligence can improve patient management at the time of a pandemic: the role of voice technology. J Med Internet Res 2021;23(5):e22959 [FREE Full text] [doi: 10.2196/22959] [Medline: 33999834]
72. Palumbo F, Crivello A, Furfari F, Girolami M, Mastropietro A, Manferdelli G, et al. 'Hi This Is NESTORE, Your Personal Assistant': design of an integrated IoT system for a personalized coach for healthy aging. Front Digit Health 2020;2:545949 [FREE Full text] [doi: 10.3389/fdgth.2020.545949] [Medline: 34713033]
73. Foehr J, Germelmann CC. Alexa, can I trust you? Exploring consumer paths to trust in smart voice-interaction technologies. J Assoc for Consum Res 2020;5(2):181-205. [doi: 10.1086/707731]
74. Gouda P, Ganni E, Chung P, Randhawa VK, Marquis-Gravel G, Avram R, et al. Feasibility of incorporating voice technology and virtual assistants in cardiovascular care and clinical trials. Curr Cardiovasc Risk Rep 2021;15(8):13 [FREE Full text] [doi: 10.1007/s12170-021-00673-9] [Medline: 34178205]
75. Sheon AR, Bolen SD, Callahan B, Shick S, Perzynski AT. Addressing disparities in diabetes management through novel approaches to encourage technology adoption and use. JMIR Diabetes 2017;2(2):e16 [FREE Full text] [doi: 10.2196/diabetes.6751] [Medline: 30291090]
76. Kowalska M, Gładyś A, Kalańska-Łukasik B, Gruz-Kwapisz M, Wojakowski W, Jadczyk T. Readiness for voice technology in patients with cardiovascular diseases: cross-sectional study. J Med Internet Res 2020;22(12):e20456 [FREE Full text] [doi: 10.2196/20456] [Medline: 33331824]
77. Coyne M, Franzese C. The Promise of Voice: Connecting Drug Delivery Through Voice-Activated Technology. East Sussex, United Kingdom: Frederick Furness Publishing Ltd; 2017.
78. Marent B, Henwood F. Digital health: a sociomaterial approach. Sociol Health Illn 2023;45(1):37-53 [FREE Full text] [doi: 10.1111/1467-9566.13538] [Medline: 36031756]
79. Ho DKH. Voice-controlled virtual assistants for the older people with visual impairment. Eye (Lond) 2018;32(1):53-54 [FREE Full text] [doi: 10.1038/eye.2017.165] [Medline: 28776586]
80. Even C, Hammann T, Heyl V, Rietz C, Wahl H, Zentel P, et al. Benefits and challenges of conversational agents in older adults : a scoping review. Z Gerontol Geriatr 2022;55(5):381-387. [doi: 10.1007/s00391-022-02085-9] [Medline: 35852588]
81. Marjanovic S, Altenhofer M, Hocking L, Chataway J, Ling T. Innovating for improved healthcare: sociotechnical and innovation systems perspectives and lessons from the NHS. Science and Public Policy 2020;47(2):1-15. [doi: 10.1093/scipol/scaa005]
82. Anastasiadou U, Alexiadis A, Polychronidou E, Votis K, Tzovaras D. A prototype educational virtual assistant for diabetes management. : IEEE; 2020 Presented at: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE); October 26-28, 2020; Cincinnati, OH, USA p. 999-1004. [doi: 10.1109/bibe50027.2020.00169]
83. Clark L, Pantidi N, Cooney O, Doyle P, Garaialde D, Edwards J, et al. What makes a good conversation? Challenges in designing truly conversational agents. : ACM; 2019 Presented at: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 4-9, 2019; Glasgow, Scotland, UK p. 1-12. [doi: 10.1145/3290605.3300705]
84. Sezgin E, Huang Y, Ramtekkar U. Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic. NPJ Digit Med 2020;3(1):122 [FREE Full text] [doi: 10.1038/s41746-020-00332-0]
85. Capasso M, Umbrello S. Responsible nudging for social good: new healthcare skills for AI-driven digital personal assistants. Med Health Care Philos 2022;25(1):11-22 [FREE Full text] [doi: 10.1007/s11019-021-10062-z] [Medline: 34822096]

Abbreviations

AI: artificial intelligence
CA: conversational agent
HIPAA: Health Insurance Portability and Accountability Act
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
SR: speech recognition

Edited by JL Raisaro; submitted 20.12.23; peer-reviewed by M Chatzimina, H Younes, H Huang; comments to author 18.04.24; revised version received 13.06.24; accepted 24.11.24; published 13.01.25.

Please cite as:
Merkel S, Schorr S
Identification of Use Cases, Target Groups, and Motivations Around Adopting Smart Speakers for Health Care and Social Care Settings: Scoping Review
JMIR AI 2025;4:e55673
URL: https://ai.jmir.org/2025/1/e55673
doi:10.2196/55673
PMID:39804689

Original Paper

Investigating the Classification of Living Kidney Donation Experiences on Reddit and Understanding the Sensitivity of ChatGPT to Prompt Engineering: Content Analysis

Joshua Nielsen[1], BS; Xiaoyu Chen[2], PhD; LaShara Davis[3], PhD; Amy Waterman[3], PhD; Monica Gentili[1], PhD

[1]Department of Industrial Engineering, JB Speed School of Engineering, University of Louisville, Louisville, KY, United States

[2]Department of Industrial and Systems Engineering, School of Engineering and Applied Sciences, University at Buffalo, Buffalo, NY, United States

[3]Patient Engagement, Diversity, and Education Division, Department of Surgery, Houston Methodist Hospital, Houston, TX, United States

Corresponding Author:
Joshua Nielsen, BS
Department of Industrial Engineering
JB Speed School of Engineering
University of Louisville
220 Eastern Parkway
Louisville, KY, 40292
United States
Phone: 1 5024891335
Email: joshua.nielsen@louisville.edu

Abstract

Background: Living kidney donation (LKD), where individuals donate one kidney while alive, plays a critical role in increasing the number of kidneys available for those experiencing kidney failure. Previous studies show that many generous people are interested in becoming living donors; however, a huge gap exists between the number of patients on the waiting list and the number of living donors yearly.

Objective: To bridge this gap, we aimed to investigate how to identify potential living donors from discussions on public social media forums so that educational interventions could later be directed to them.

Methods: Using Reddit forums as an example, this study described the classification of Reddit content shared about LKD into three classes: (1) present (presently dealing with LKD personally), (2) past (dealt with LKD personally in the past), and (3) other (LKD general comments). An evaluation was conducted comparing a fine-tuned distilled version of the Bidirectional Encoder Representations from Transformers (BERT) model with inference using GPT-3.5 (ChatGPT). To systematically evaluate ChatGPT's sensitivity to distinguishing between the 3 prompt categories, we used a comprehensive prompt engineering strategy encompassing a full factorial analysis in 48 runs. A novel prompt engineering approach, dialogue until classification consensus, was introduced to simulate a deliberation between 2 domain experts until a consensus on classification was achieved.

Results: BERT and GPT-3.5 exhibited classification accuracies of approximately 75% and 78%, respectively. Recognizing the inherent ambiguity between classes, a post hoc analysis of incorrect predictions revealed sensible reasoning and acceptable errors in the predictive models. Considering these acceptable mismatched predictions, the accuracy improved to 89.3% for BERT and 90.7% for GPT-3.5.

Conclusions: Large language models, such as GPT-3.5, are highly capable of detecting and categorizing LKD-targeted content on social media forums. They are sensitive to instructions, and the introduced dialogue until classification consensus method exhibited superior performance over stand-alone reasoning, highlighting the merit in advancing prompt engineering methodologies. The models can produce appropriate contextual reasoning, even when final conclusions differ from their human counterparts.

(JMIR AI 2025;4:e57319) doi:10.2196/57319

KEYWORDS

prompt engineering; generative artificial intelligence; kidney donation; transplant; living donor

Introduction

Background

Kidney transplantation is the gold standard treatment for patients with end-stage renal disease [1] and can be much more cost-effective than dialysis [2]. Record numbers of transplants have taken place in recent years, but a shortage of donors continues to exist despite the recent increase [3]. Currently, the median wait time for a transplant is approximately 4 years in the United States, and approximately 5000 patients die every year while being on the transplant waiting list [4]. Living donor kidney transplantation (LDKT) generally provides better outcomes than deceased donor transplants but requires that a potential living donor be made aware that they can donate to a specific patient with end-stage renal disease and offer to do so. Racial or ethnic minorities and patients of lower socioeconomic status are less likely to pursue and have living donors donate on their behalf [5,6].

National attitudes about LDKT are generally positive, although many do not know what a living donor undergoes when donating a kidney [7-10]. Recommendations to increase the living donor pool include reaching out more broadly to locate generous individuals motivated by social good to engage more individuals in considering living donation [11]. In addition, research suggests that disseminating education and information about living donation to broader audiences, beyond transplant centers, might increase the numbers of potential donors and recipients pursuing living donation [12,13]. However, identifying individuals dealing with kidney disease and considering whether to pursue LDKT or donate kidneys in their own lives can be difficult, especially when they have not started medical evaluation at a transplant center.

Locating individuals through social media forums discussing living kidney donation (LKD), such as those on Reddit or Twitter (the work herein was done before the platform being rebranded as X), maybe a way to identify individuals who are actively deciding whether to pursue LDKT or LKD outside of transplant centers [14]. While there are many different types of questions and comments related to LKD shared on the web, some people share their personal experiences and even invite people to “ask me anything.” These findings motivated our main hypothesis that potential living donors can be identified from social media communities engaged in general discussions about LKD. In addition, understanding the personal experiences shared on these platforms can provide valuable insights into potential donors’ needs and decision-making, enabling education and media campaigns to be better tailored for them.

The large volume and high complexity of unstructured natural language require an effective and efficient method that can automate the identification of people sharing personal experiences with LKD. Fortunately, recent advances in natural language processing (NLP), particularly the transformer mechanism [15-19], enable the automatic understanding of personal experiences that were shared on the web social platforms. This study aimed to evaluate the transformer-based techniques to categorize these experiences on Reddit (Reddit, Inc). Specifically, we aimed to evaluate and compare (1) the one-shot classification model Bidirectional Encoder Representations from Transformers (BERT) [19], which required that we fine-tune the model using 1268 well-labeled samples, and (2) the zero-shot classification model ChatGPT (OpenAI), which required no fine-tuning for classification purposes. Comprehensive discussions on transformer-based models can be found in the study by Acheampong et al [20]. Much has been written about the capabilities and limitations of ChatGPT specifically [21]; however, we investigated the importance of prompt engineering when interfacing with it and other generative models applied to the field of organ donation for the first time.

Overview of Prompt Engineering

Prompt engineering has been defined as “the means by which LLMs are programmed via prompts” [22]. Reynolds and McDonell [23] framed the objective of prompt engineering as a discipline that seeks to answer the question, “What prompt will result in the intended behavior and *only* the intended behavior?” Historically, the best practice has been to give a small number of examples of how the task is to be done, known as few-shot prompting. Ray [21] suggested that for large language models (LLMs), few-shot prompting is better thought of as “locating an already-learned task rather than meta-learning.” The implication is that the LLMs are large and robust enough that the models are inherently capable of completing NLP tasks, but their scale of capability may require using examples to “activate” the right parameters that will carry out the desired task in the prescribed manner.

However, this flexibility should also be understood as having dangers because LLMs can be “jailbroken.” Jailbreaking LLMs is the practice of using prompt engineering to work around the boundaries imposed by the developers, such as OpenAI [24]. The practice of “red-teaming” is used by developers to identify weaknesses in the desired boundaries and adjust the model so that it is more defensible against previous vulnerabilities [25,26]. What is simultaneously exciting and problematic about this is that many techniques used to jailbreak LLMs are the same as those used for their most helpful, intended uses, that is, many of the same methods that allow us to get the best performance from an LLM can be the same ones that are used to bypass the safeguards. Table 1 provides an overview of prompt engineering methods derived primarily from the study by White et al [22].

Table 1. Overview of prompt engineering methods proposed by White et al [22].

Method	Purpose	Example prompts for LKD[a]
Few-shot prompting	Provide examples that illustrate how the task is to be completed	"Here is an example of a risk analysis from a living kidney donation scenario: [EXAMPLE]. Now, please provide a risk analysis for the following scenario."
Meta-language creation	Create a shorthand notation, abbreviated language, or set of standard rules	"For this conversation, 'LKD' refers to living kidney donation, 'DT' refers to donor testing and 'RC' refers to recipient compatibility. Using this shorthand, describe the typical process of LKD."
Flipped interaction	The LLM[b] will ask questions to obtain the information	"I'm working on an algorithm to match donors with recipients in living kidney donation. What information do you need from me to help design this algorithm?"
Persona	Assign a persona to the LLM, usually that of an expert	"Pretend you are a leading surgeon specializing in living kidney donation. Provide your expert opinion on the latest surgical techniques."
Prompt refinement	Ensure that the LLM suggests better or more refined prompts	"I need to write code to analyze the success rates of different kidney matching algorithms. Could you suggest a more refined question or specific details you need to assist me?"
Alternative approaches	Ensure that the LLM offers alternative ways of accomplishing the task	"Describe three different methods for assessing donor-recipient compatibility in living kidney donation."
Cognitive verifier	Subdivide a question into additional questions for a better answer	"To understand the ethical considerations in living kidney donation, what additional questions should I ask you to provide a comprehensive analysis?"
Fact checklist	Mitigate model hallucination by listing the facts	"After explaining the current trends in living kidney donation, list the facts or data sources you used in your response."
Template	Ensure that the LLM's output follows a precise template	"Please answer in the following format: 'Living kidney donation is beneficial because [REASON 1], [REASON 2], and [REASON 3]'."
Gameplay	Create a game around a given topic	"Let's play a matching game. I will describe a recipient, and you suggest a suitable donor from the provided pool based on living kidney donation criteria."
Reflection (chain of thought [25])	Explain the rationale behind the given answers	"Explain the process of donor selection in living kidney donation in a step-by-step manner, detailing the reasoning behind each step."
Refusal breaker	Help users rephrase a question when they are refused an answer	"If you cannot provide personal patient data in living kidney donation, please guide me on how to rephrase my questions to obtain general information."
Context manager	Enable users to specify or remove context	"When discussing living kidney donation statistics, please consider only data from the last five years in the European region."
Recipe	Provide a sequence of steps given some partially provided ingredients	"I have patient medical records, compatibility testing results, and surgical schedules. Provide a sequence of steps to create an optimal living kidney donation matching algorithm."

[a]LKD: living kidney donation.

[b]LLM: large language model.

Reflection and chain of thought reasoning, in particular, have garnered much attention due to their powerful results, creating what is already becoming a niche corner of research [27,28]. At the time of writing this paper and to the best of our knowledge, the 2 most recent and powerful of these improvements are the methods known as self-consistency [29] and the tree of thoughts [30]. The former uses majority voting from multiple replications, and the latter takes an ensemble approach to the chain of thought reasoning and allows LLMs to consider multiple different reasoning paths and to perform self-evaluation on choices. Other methods naturally exist beyond what is contained in this study because of the unbounded human imagination, which makes the domain of prompt engineering quite an exciting frontier. Interested readers may find the website [31] to be a useful resource, with new relevant articles being added to its repository regularly.

While prompt engineering in the context of LKD has not yet entered the literature, some work has emerged in the context of health care. Prompt engineering and generative artificial intelligence broadly are of particular interest in the medical domain as the generation of health information is still of unknown quality. A few researchers have emphasized the importance of medical professionals using LLMs skillfully and in a way that produces reliable information [32,33]. It has been shown that the reliability of GPT-4 (OpenAI) is inconsistent when answering medical questions, and the authors call for prompt engineering techniques to improve its performance [34]. Similarly, other authors have experimented with ChatGPT on calculation-based United States Medical Licensing Examination questions using 3 different prompting strategies, although they found that the prompt itself had only a small effect on answer accuracy [35]. Other research examined using prompt

engineering in generating health messages [36] and even medical image segmentation [37].

Social Media and LKD

Recent years have witnessed a burgeoning interest in studying dialogue on social media regarding important health care issues, such as vaccination [38] and LKD. Henderson [39] highlighted the use of platforms such as Facebook and Twitter to identify potential living donors while noting that formal research efforts are in their early stages. Analyzing social media content, including organ donation posts on the Chinese social media site Weibo, has unearthed key themes such as "organ donation behaviors," "statistical descriptions of organ donation," and "meaningfulness of donation" [40]. In one study, a notable 53% of potential living donors who self-referred for donor evaluation reported that they learned about a patient's need for a donor on social media [41,42], while specialized tools such as the "DONOR" app have enabled expansion of social media marketing about living donation between potential donors and patients with kidney diseases [43]. Research efforts include measuring organ donation awareness through Twitter digital markers [44], surveying readiness of patients who are undergoing a transplant to use social media for education [45], and using Twitter for living donor profile classification [46].

Interventions to increase living donation have used mobile health technologies to manage donor follow-up [47], delivered targeted advertising to specific ethnic groups [48,49], and assessed organ donation awareness across the United States using Twitter data [50]. Best practices for promoting LKD through social media, such as delivering content to specific community demographics in targeted and interactive modes, have been proposed [51]; live transplant broadcasts on Twitter have occurred [52]; and the analysis of public Facebook pages of potential living donors [53] has enhanced insights into donor identification and donation interest. Recent studies highlighted the importance of tailored messaging over generic communication for better audience engagement [54,55].

These investigations underscore social media's potential in augmenting donation awareness and facilitation, emphasizing the necessity for robust methods to discern and support individuals disseminating LKD-related content. A recent study by Garcia Valencia et al [56] has shown that ChatGPT can simplify medical information, making it easier to read and understand by many diverse groups. This can be a vital aid for promoting fairness in access to donation information from official sources. However, with the availability of *public* dialogue in forums also comes the need to thematically understand it. There is variation in both the content being shared and the user sharing it. The growing body of research demonstrates the potential of social media to impact awareness, intention to donate, and the facilitation of living kidney transplants. Therefore, it is necessary to have reliable methods whereby people who explicitly create and share content related to LKD can be automatically identified and understood for appropriate education and support. With this background, our research seeks to assess whether a classification system can be devised to discern individuals at varying stages of decision-making about becoming a living kidney donor. It also explores which of the contemporary NLP models are most apt for automating this classification, namely a fine-tuned distilled version of the BERT (DistilBERT) model (hereafter referred to as BERT for simplicity, unless greater specificity is merited) or ChatGPT. Furthermore, regarding ChatGPT, it examines how prompt engineering—namely, making adjustments to model instructions about the reasoning approach, examples, temperature, and class descriptions— influences its predictive efficacy for this application.

By answering these research questions, this study aimed to build a foundation for a sophisticated classification system in which it is possible to automatically categorize large amounts of social media communication about living donations using these tools. The study also aspires to gain a more in-depth insight into how individuals communicate and express themselves regarding LKD on various social media platforms. Using cutting-edge NLP technologies, our goal is to develop a streamlined, automated process for pinpointing curious, motivated potential donors who have not yet presented to the transplant center so that educational interventions could later be directed to them.

Methods

Data Labeling, Preparation, and Quality Assurance

We used a dataset of 2689 Reddit posts related to LKD from our previous work [14], which were published between January 2010 and April 2021. We also collected 603 Reddit posts from April 2021 to April 2023, for a combined total of 3292 posts from 2591 users. We scraped the posts with the open-source tool pushshift.io using keywords related to LKD, such as "kidney donor," "kidney transplant," "kidney donated," "kidney donate," "kidney years ago," "kidney need," "kidney stranger," and "kidney willing donate." Other search terms could have been included; however, as presented in Table 2, a considerable portion of collected data were not related to personal experiences, and we concluded that additional search terms would primarily expand the noise and add little value.

Table 2. Distribution and description of Reddit (Reddit, Inc) classes.

Merged class categories and class categories	Description	Example post
Present (n=540, 26.9%)		
Present direct (n=363, 21.5%)	The user has *current firsthand experience* with something personally related to kidney disease, kidney failure, living kidney donation, or transplantation (eg, the user with kidney disease or kidney failure, is on dialysis, is seeking a kidney, is exploring donation, or is undergoing evaluation for donation or transplantation).	"A friend of mine is in need of a kidney. My first instinct is to offer one of mine. I have Googled and read LOTS of info. What would you do? Have you donated a kidney? What am I missing?"
Present indirect (n=177, 5.4%)	The user has *current secondhand experience* related to living kidney transplantation (eg, they *know someone* who is currently experiencing kidney failure, on dialysis, seeking a kidney, or preparing to donate a kidney).	"I need help finding a kidney for my dad."
Past (n=222, 6.8%)		
Past direct (n=168, 5.1%)	The user has *past firsthand experience* related to living kidney transplantation (eg, kidney failure, dialysis, kidney recipient or donor).	"Eight years ago today, I donated a kidney to a friend. Ask me anything."
Past indirect (n=58, 1.8%)	The user has *past secondhand experience* related to living kidney transplantation (eg, they *know someone* who experienced kidney failure, was on dialysis, received a kidney, donated a kidney, underwent evaluation for donation, or participated in the donation process (perhaps in a supporting role).	"Picture of my dad and the woman who donated a kidney to save his life."
Other (n=2530, 76.8%)		
General commentary or hypothetical (n=159, 4.8%)	The user is giving a *general opinion* on the topic, asking a *hypothetical question*, or contributing to discussion about an *imagined scenario*.	"If you donate a kidney, then later your only one starts to fail, would you be put on a higher priority?"
News or noise (n=2371, 72%)	The user is either sharing a *news article or headline* related to kidney donation that may be pertinent but *not personal*, or it is *simply irrelevant*.	"A man donated his kidney to his wife of 51 years after finding out he's her perfect match."

We selected Reddit as our data source because it provided the greatest portion of comments that were related to personal experiences rather than discussions of policies and sharing news stories. Reddit was the only place where we found posts from actual living donors inviting people to an "ask me anything" session, sparking highly personal discussions [14].

Under the guidance of LKD domain experts, after reviewing 100 example posts, we created 2 class sets, one with 6 classes (class categories) and the other with 3 classes (merged categories), to automate the process of identifying firsthand experiences with living donation (Table 2). These classes were iteratively defined and improved through multiple discussions with a team of 6 people who performed the manual annotation. Certain posts had sufficient ambiguity to make an explicit ruling impossible. For example, it was not always clear what constituted the boundary between a past and present experience (eg, how much time should have passed since the transplant?) or whether the general transplant mentioned in a post came from a living or deceased donor. Furthermore, long and verbose posts with brief mentions of personal experiences with donation posed a challenge because the brief (although important) mentions of LKD were easy to miss. Individual annotators were found to exhibit varying classification tendencies or use their own "rules of thumb" to expedite the often tedious process.

The granularity between these 6 fine-grained classes proved quite difficult for the models to correctly capture during initial experiments (resulting in accuracies <50%), so the posts were consolidated into the 3 coarse-grained categories: present (n=540, 42.59% of posts), past (n=222, 17.51% of posts), and other (n=506, 39.91% of posts randomly sampled from news or noise and general commentary or hypothetical categories) for 1268 samples that were used for training the BERT model. A randomly selected subset of 100 from each of the 3 classes was used for prompting with ChatGPT. The decision was made to aggregate general commentary and hypothetical posts with news or noise to ensure a more precise focus on personal experiences.

Acknowledging the potential data quality risks [57], we meticulously evaluated incorrect predictions from both BERT and ChatGPT after the analysis. The incorrectly predicted samples were tagged as either acceptable errors (reasonable, if not perfectly aligned predictions), unacceptable errors (flawed or evidently incorrect reasoning), more accurate than the original human label, or instances where both human and model erred. We later reported these using the notation of *LLM human*, *LLM<human*, *LLM>human*, and *both error*, respectively, for both models.

Ethical Considerations

This study was granted an exemption from The University of Louisville Institutional Review Board (review number 22.0458). While there could be ethical concerns about consent and storage of health-related data, every Reddit user is entirely anonymous, ensuring that nothing we find can be directly traced to an individual. In addition, the comments and posts themselves are all very public; some websites may have minimal requirements, such as logging in or being a member of a "closed" group before the content can be observed; however, this is not the case for any of the data we collected. For data sources where such anonymity is not guaranteed, it is imperative to ensure that users consent to the study of their created content and that any identifying information be removed or obscured.

Modeling

We compared 2 transformer-based models for our classification task: a fine-tuned BERT model and a prompt-engineered ChatGPT model. We used the 3.5 Turbo version of ChatGPT via the OpenAI application programming interface and conducted a full factorial analysis of various prompt components to identify the best features. The DistilBERT model was fine-tuned from a pretrained Hugging Face (Hugging Face, Inc) model. Furthermore, we noted that many new models have emerged, both proprietary and open source, after our experiments were completed. Post hoc experiments indicate that our findings are consistent with newer models.

BERT Analysis

The DistilBert tokenizer from Hugging Face was used to tokenize the text data from Reddit, and both input IDs and attention masks were generated to structure the text inputs for the model. A custom model was designed around DistilBERT. The architecture included the pretrained DistilBERT model, followed by 3 fully connected layers with 768, 256, and 128 units, respectively. These were followed by an output layer with 3 units corresponding to the number of classes. Batch normalization and rectified linear unit activation functions were applied, and dropout was set at 10%.

The focal loss was used as the loss function, which is designed to address the class imbalance by downweighting the loss assigned to well-classified examples [58]. It was parameterized with an α factor for controlling the weight and a γ factor for focusing on hard examples. The model was trained using the AdamW optimizer [59], with the learning rate and weight decay optimized by the open-source Optuna hyperparameter tuning library. The dataset was split into training and validation sets using stratified 5-fold cross-validation, with class weights computed to manage class imbalance, and the model was trained for 3 epochs, following the recommended fine-tuning procedures [19]. The metrics used for validation are defined subsequently.

Accuracy is the ratio of correctly predicted instances to the total instances.

Precision is the ratio of correctly predicted positive observations to the total predicted positives.

Recall (sensitivity) is the ratio of correctly predicted positive observations to all observations in actual class.

F_1-score is the harmonic mean of precision and recall.

In equations 1 to 4, *TP*, *TN*, *FP*, and *FN* are the numbers of true positive, true negative, false positive, and false negative values, respectively.

The Optuna library was used to perform hyperparameter optimization, which uses a Bayesian optimization method known as the Tree-structured Parzen Estimator [60]. A search space was defined for the learning rate (ranging from 0.00003 to 0.0003) and weight decay (ranging from 0.0001-0.001). A total of 100 trials were conducted to find the best set of hyperparameters based on the F_1-score.

Dialogue Until Classification Consensus

We introduced a text classification tool for LLMs termed "dialogue until classification consensus" (DUCC). Given the absence of a formal taxonomy for prompt engineering methods, we aligned DUCC's presentation with the pattern widely adopted in software development, which includes a name and classification, intent and context, motivation, structure and key ideas, example implementation, and consequences (Textbox 1). White et al [22] constructed the following categories of prompting patterns: input semantics, output customization, error identification, prompt improvement, interaction, and context control.

Textbox 1. Prompting patterns for "dialogue until classification consensus" (DUCC).

Name and classification

DUCC primarily falls under output customization, although it shares elements from other pattern categories, notably error identification and interaction.

Intent and context

DUCC assigns a persona of at least 2 domain experts to the large language model, instructing them to discuss a text sample until a consensus on its classification or answer selection is reached from a set of options. This setup aims to automate explicit reasoning and reflection through a simulated dialogue, expecting to resemble the effects of distribution-oriented methods, such as self-consistency, without requiring multiple sample replications.

Motivation

Complex classification tasks, especially within niche domains, such as personal living kidney donation experiences, often present labeling challenges. DUCC simulates expert discussions for decision-making while aiming to standardize output formats for classification tasks.

Structure and key ideas

Experts 1 and 2, specialized in [DOMAIN], are to discuss the text sample until an agreed classification or answer is reached.

The final label should be clear with no disagreements, formatted as: "classification: Label."

Additional identities or traits can be attributed to the experts to infuse specific perspectives into the discussion. We have observed that unless a singular label selection is emphasized, the model might assign multiple labels in challenging scenarios.

Example implementation

"Expert 1 and Expert 2, you are both experts in living kidney donation, and you've been tasked with analyzing and classifying a Reddit post that should be related to living kidney donation. You should discuss the post until you come to an agreement for a single classification. If the post is not related to living kidney donation, it needs to be labeled 'Other'. The classifications are defined as follows:

- Present: The user is describing a current or ongoing personal experience with living kidney donation
- Past: The user is describing a past personal experience with living kidney donation.
- Other: The user isn't discussing a personal experience with living kidney donation or isn't discussing living kidney donation at all.

Discuss until you reach a consensus, showing your reasoning. The final label should be clear, and there should be no disagreement. Output your agreed label in this format: {'classification': 'your agreed label'}.

Here's an example of how this should be done:

- Post: 'Are you a kidney donor? How was the recovery process and how are you doing now?'
- Expert 1: 'I think the appropriate label is Present, because the user is asking questions and seems to want information to help them with a current decision about living kidney donation.'
- Expert 2: 'I think the appropriate label is Past because the user wants to know about past personal experiences from others.'
- Expert 1: 'I see your point about bringing up the past, but since we are interested in assigning a label to the user who wrote the post, we should keep our focus on the author's perspective. If we knew what the replies were, we could label those users as Past, but we are only looking at this user for now.'
- Expert 2: 'You're correct, we should be focused on this user rather than possible answers from others. Even though there are elements of both, we have to pick one and only one label, so let's go with Present.'
- Final Label: "'classification': 'Present.'"

Consequences

DUCC prompts large language models to reason through multiple perspectives, ensuring a singular, consistently formatted label, simplifying extraction. The example implementation is crucial as it demonstrates the desired dialogue structure, aiding the model in handling nuanced classifications. However, DUCC may exhibit biases when numerous classes are present, potentially leaning toward the exemplified label. To mitigate token use, especially in lengthy examples, using DUCC when defining the system instead of individual prompts is advisable. For instance, in the OpenAI application programming interface, modifying the "content" section of the "system" role with the entire provided example instead of the default content can better define the system's nature.

Sensitivity Analysis of Prompting

Overview

For our experimentation using ChatGPT to categorize personal experiences, we conducted a study applying a full factorial design with 4 factors (summarized subsequently), which resulted in 48 experimental runs. We must first acknowledge that the nature of prompting is such that there were an infinite number of ways we could write the prompt and parameters that could be chosen. It is well known that examples that illustrate the solutions can influence performance (known as "few-shot" prompting) [61], so we examined the number of examples and the type of examples that might produce bias as well as the parameters provided subsequently.

Use of the DUCC Method (2 Settings)

In addition to the DUCC method described earlier, the alternative was to prompt a single expert to make a classification decision, with the instruction to "Examine the evidence for each class option step by step. The final label should be clear." In this case, the model attempts to identify any evidence that suggests the sample should be assigned to each class and weighs the evidence to draw a conclusion.

Number of Examples Used (4 Settings)

We selected either 1 example or 3 examples. For 3 examples, 1 example was used for each class (present, past, and other). For the single example setting, we performed an experiment with each class once to evaluate whether it produced a bias in the predicted class.

Definition of "Past" (2 Settings)

Observing a tendency for underprediction in the "Past" label, we considered 2 definitions for the class. The first was a short and concise definition: "The user is describing a past personal experience with living kidney donation." The second was a longer, more descriptive definition: "The user is referring to a past personal experience with LKD. This may be presented in the context of a present tense story, but if the event of LKD was lived previously, the post should be labeled past."

Temperature Settings (3 Settings)

Experimentation spanned temperature values of 0, 0.15, and 0.3, investigating the tradeoff between output variability and consistency. The settings were guided by OpenAI documentation, emphasizing lower values for consistency and higher values for diversifying outputs [62].

Given the cost implications of OpenAI application programming interface calls, an initial assessment was carried out to determine the necessity for replicating each setting. We performed 30 replications of a fixed parameter setting and found no substantial effect within replications for any metric. Thus, the experimentation proceeded with a singular sample for each parameter setting.

Results

Overview

In this section, we present the results of the BERT model first and then the results of ChatGPT. We present the performance metrics, confusion matrices, and assessment of incorrect predictions. For ChatGPT, we also present the results of an ANOVA on the various factors used in the experimentation.

BERT Results

In >100 trials, the best BERT model performed with an accuracy of 75.1% and an F_1-score of 78.2% on the validation data during training. The best parameters were a learning rate of 0.000131687 and a weight decay of 0.000791. The confusion matrix for the predictions on the test data is presented in Figure 1, showing reasonably good performance but with a tendency to erroneously predict the Other label on both past and present labels.

The classification report provided in Table 3 shows that the BERT model significantly underpredicts past labels, partly due to the smaller sample size, and also because of the ambiguity that can arise when a reference to a past experience is nested within an ongoing story.

Figure 1. Confusion matrix for the best Bidirectional Encoder Representations from Transformers model.

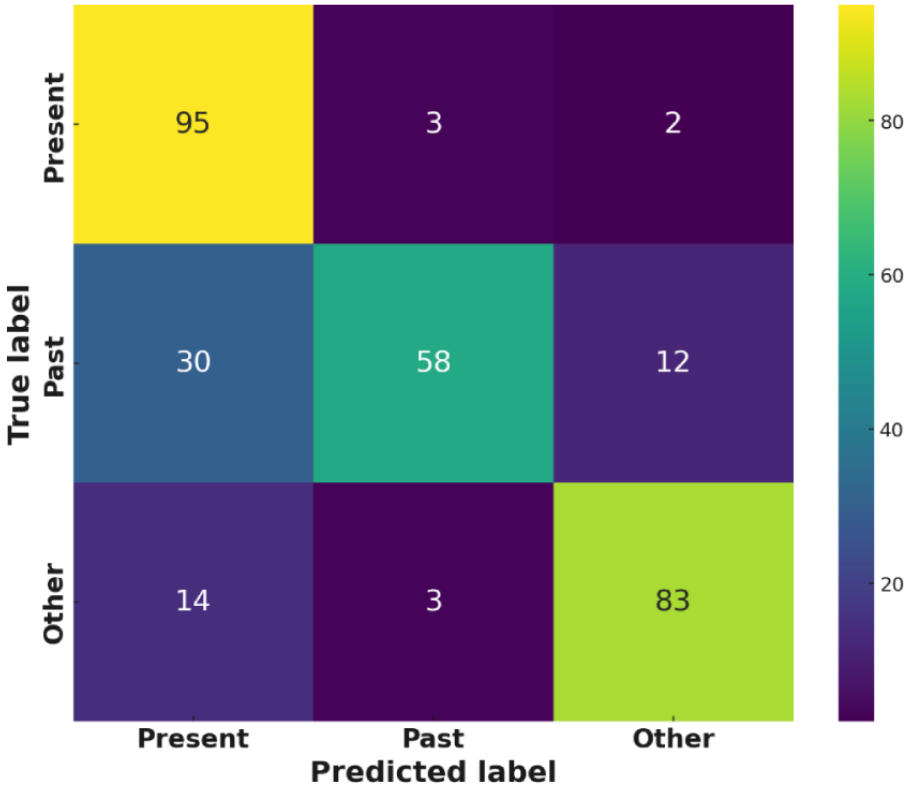


Table 3. Classification report.

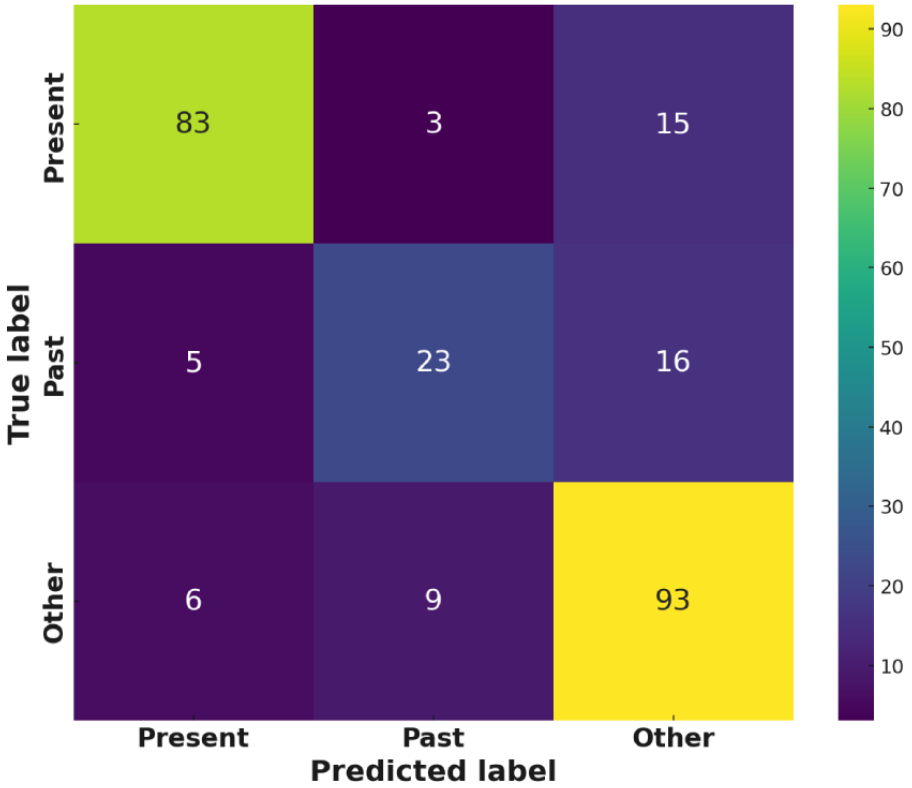
	Precision	Recall	F_1-score	Support
Present	0.88	0.82	0.85	101
Past	0.66	0.52	0.58	44
Other	0.75	0.86	0.80	108
Weighted average	0.79	0.79	0.78	253

ChatGPT Results

The best ChatGPT prompt produced an accuracy and F_1-score of 78.67% and 78.17%, respectively (surprisingly, this F_1-score is identical to that of BERT). This was achieved using the DUCC method, a single example of a present class post, a temperature of 0, and the shorter definition of the past class (refer to the Dialogue Until Classification Consensus section). Full experimentation results are provided in the Multimedia Appendix 1. The next 3 columns show the percentage of predictions for that class, and the remaining 3 columns show the evaluation metrics.

The confusion matrix for ChatGPT performance is presented in Figure 2, which shows again that past class samples were underpredicted and that both other and past class samples were overpredicted to be present class, suggesting a bias toward present classifications.

Figure 2. Confusion matrix for the best ChatGPT prompt.



The results of the ANOVA are presented in Table 4, which shows that the number and type of examples used is the most significant factor, followed by the method. We observe that the examples and method factors were the only statistically significant factors.

Given that there were 3 df within the examples setting, we sought to better understand the difference between the example settings using a Tukey test, with results provided in Table 5. We observed that when our example belonged to the "past" class the model performed better than when the example came from the "other" class. But using an example from the "past" class resulted in poorer performance compared to using 3 examples (one from each class) and using an example from the "present" class. Interestingly, the "past" sample was underpredicted in every setting except when using 3 examples and the evidence method. Interestingly, samples belonging to the "past" class were underpredicted in every setting except when using 3 examples and the evidence method. Although this setting (3 examples; evidence method) does not demonstrate the same underprediction bias as other settings, it does not give better accuracy overall.

Table 4. ANOVA results.

Factor	Sum of squares	*F* test (*df*)	*P* value
Category (examples)	0.068615	27.659884 (3, 40)	<.001
Category (method)	0.006466	7.819650 (1, 40)	.008
Category (temp)	0.000024	0.014557 (2, 40)	.99
Category (past)	0.000032	0.039292 (1, 40)	.84
Residual	0.033076	—[a]	—

[a]Not applicable.

Table 5. Multiple comparisons of means using the Tukey honestly significant difference test. The family-wise error rate is 0.05.

Group 1	Group 2	Mean difference	*P* value	Lower limit	Upper limit	Reject
1, other	1, past	–0.0875	<.001	–0.1202	–0.0548	True
1, other	1, present	0.0078	.92	–0.0249	0.0405	False
1, other	3	–0.0017	.99	–0.0344	0.031	False
1, past	1, present	0.0953	<.001	0.0626	0.128	True
1, past	3	0.0858	<.001	0.0531	0.1185	True
1, present	3	–0.0094	.87	–0.0421	0.0233	False

Discussion

Principal Findings

Our experimentation has found that BERT and ChatGPT perform comparably for the classification of different living kidney donor experiences. Because BERT is completely dependent on the available training data, ChatGPT can be used with a somewhat higher degree of precision via prompt engineering, as shown by our use of the novel DUCC method. Our full factorial experimentation identified the best settings to use for our engineered prompt. In this section, we will discuss the predictions that were made incorrectly and consider future work and ethical considerations.

Examination of Incorrect Predictions

As noted in the Data Labeling, Preparation, and Quality Assurance section, there is an inherent risk of data quality that arises from the dataset in question. Unlike standardized benchmarks, which often have explicit “ground truth” labels, our task is fraught with nuance. Despite our extensive efforts to ensure data quality, the given label is not always clear. As such, we have provided a more detailed examination of the instances where the models made predictions that diverged from the given labels.

BERT and GPT-3.5 produced 21.3% (54/253) and 21.3% (64/300) incorrect predictions, respectively. It should be recalled that the difference in the denominator values is because BERT requires a split test set, whereas, with GPT-3.5, we can use a larger inference-only set. We assessed the quality of these incorrect predictions not only to see how “close” they were to the mark but also to determine whether any human errors had been made in labeling the incorrect predictions. As provided in Table 6 for BERT, we observe that 27 prompts were incorrectly labeled either because of an acceptable error where a clear prediction is difficult to make (perhaps due to the ambiguity of what constitutes the difference between the past and present samples) or where BERT made a better prediction than the original human label. Treating these 27 predictions as being acceptable or correct brings the total number of correct predictions from 199 (78.7%) of 253 to 226 (89.3%) of 253, which elevates the predictive accuracy considerably to 89.3%. In these tables, examples are written “as they are” from the original posts, including typos and terminology that may be unique to Reddit.

Table 6. Analysis of incorrect predictions from Bidirectional Encoder Representations from Transformers (BERT; n=54).

Error type	Incorrect predictions, n (%)	Example post	Reason
Unacceptable error (BERT<human)	22 (41)	"Required testing to be a living Kidney donor where I live - these are the tests I took before becoming a living kidney donor almost 2 yrs ago everything has gone great for me and the recipient happy to answer any questions."	BERT predicted the "other" label, but the user clearly states that he or she was a previous living donor.
Acceptable error (BERT human)	12 (22)	"Hey Mum, it's been a year since what was supposed to be a life changing kidney transplant that took a turn for the worst. I love you so much and think about you every day xxx"	BERT predicted the "other" label, which could be appropriate if it was a deceased donor transplant. We predicted the "past" label.
Human error (BERT>human)	15 (27)	"Me 26F with my Dad 58 he needs a kidney and I feel pressured to donate one. [removed]"	We predicted the "other" label because of the (removed) tag at the end of the post, which commonly appears in unusable posts. BERT predicted the "present" label, which is the more appropriate label.
Both erred	5 (9)	"I used to like her but I found out that she did not even acknowledge her kidney donor... Just referring to her as a person I know it seems pretty ungrateful [removed]"	This is someone's opinion about a celebrity who famously received a kidney transplant from her friend. It is not a personal experience at all, but the human label was "present," and the BERT label was "past."

From our analysis of the incorrect predictions on GPT-3.5 (Table 7), we observed that 26 (40%) of the 64 errors were acceptable.

As mentioned earlier, we had previously observed that many "past" posts were labeled as "present" because many of the posts were in a present tense context. The best setting used the shorter definition of past, which does not teach the model to treat past experiences nested in present accounts as the past class, so this is to be expected. Anytime both the human and predicted labels were wrong, the post was almost always ambiguous regarding whether it was about living or deceased donation. The experiences being described could have been a living donation, but there is not enough information to determine that for certain.

Regarding BERT, we may allow ourselves to consider the 26 acceptable errors and 10 human errors as being correctly predicted, changing the total number of correct predictions from 236 (78.7%) of 300 to 272 (90.7%) of 300 for an "actual" predictive accuracy of 90.7%. While still imperfect, this shows considerable reliability when using these methods on nuanced language tasks.

The implications of this examination are threefold: (1) sometimes human annotations go wrong, even with clear instructions; (2) these powerful models are capable of correctly catching things that humans miss (due to decision fatigue or similar cognitive difficulties); and (3) the models can be largely trusted to give sensible reasoning, even if the final conclusions differ from that of a human counterpart.

Table 7. Analysis of incorrect predictions from ChatGPT (n=64).

Error type	Incorrect predictions, n (%)	Example post	Reason
Unacceptable error (ChatGPT<human)	21 (33)	"relationships My (36F) estranged sister (43F) donated a kidney to me. I just heard that she died (for a different reason). I'm very confused. [removed]"	The simulated experts reasoned that the focus of the post was on grief rather than LKD[a] and labeled it as "other." The human label was given as "past" because the user mentions a sister who donated her kidney some time ago.
Acceptable error (ChatGPT human)	26 (41)	"Successfully donated a kidney to my sister whos been fighting Lupus."	This could be easily interpreted as either a "present" (ChatGPT) or a "past" (human) label, given that there is no explicit reference to time. It could go either way, but it is still clearly related to a personal experience with LKD.
Human error (ChatGPT > human)	10 (16)	"I (30F) had heart and kidney transplant. Ask Me Anything (AMA)."	The simulated experts concluded that this should be labeled "other" when the human label had been given as "past." ChatGPT made a more correct conclusion because this may have been from a deceased donor rather than a living donor. We would need more information to be certain, so it should be an "other" label.
Both erred	7 (11)	"I am A double kidney transplant recipient! AMA! I am a 28 year old white male, I've had two renal transplants over the course of my lifetime. I've been on dialysis. I've been in and out of hospital my entire life. I think it's interesting, but there's only one way to find out! Ask Me Anything."	The human-given label for this was "past" because of the previous transplant experiences, and the reasoning provided by ChatGPT concluded that the label should be "present" because the user mentions dialysis and being in and out of the hospital. Both were incorrect because there is not enough evidence that either of the transplants was from living donors, and thus, it should be labeled "other."

[a]LKD: living kidney donation.

Limitations and Future Work

BERT and ChatGPT have both proven effective in classifying personal accounts of LKD on platforms such as Reddit, achieving approximately 80% accuracy, which increases to about 90% when considering acceptable errors, marking a step forward in using web-based data for LKD research. These models could potentially automate the screening of new content for further scrutiny, thereby aiding donor support initiatives, particularly in education and community outreach. Despite the promising results, the complexity of the subject matter complicates the task of making perfect predictions. Our initial attempts to use fine-grained classifications led to suboptimal results, requiring us to use coarse-grained categories. Regarding costs, BERT's open-source nature and the flexibility to fine-tune make it an appealing choice. In contrast, ChatGPT excels in providing understandable reasoning for its decisions.

A review of errors indicated that ChatGPT generally understood the context well, although there were instances where the reasoning was off the mark, highlighting the importance of clear, prompt instructions. Interestingly, there were instances where the LLMs' reasoning surpassed ours, especially in delineating the "past" and "present" boundary, thereby suggesting a potential for iterative prompt enhancements informed by LLM reasoning. However, the quest for prompt optimization (or "promptization," if you will) may present an unending journey, as the allure of "just one more experiment" to elevate performance is always present. Drawing a line on performance as "good enough" is crucial, which may be attained through automated processes, as explored in some recent and exciting studies [63-69]. Future work will leverage these powerful new methodologies to both improve performance on our coarse-grained 3-class schema as well as achieve superior performance on the fine-grained 6-class schema that was unattainable with the present methods.

The performance of both models is significantly constrained by the size of the available data. While thousands of Reddit posts related to LKD are accessible, only a fraction pertains to personal experiences. The performance consistency across different data folds for BERT and across different sample sizes for ChatGPT highlights the need for larger datasets to better gauge each model's robustness.

A core challenge lies in the task's inherent demand for a singular label, which often oversimplifies the nuanced narratives in internet posts. Future endeavors could explore more elaborate information extraction techniques, leveraging LLMs such as ChatGPT to answer multiple queries or even construct knowledge graphs per post. Although ensuring uniform and usable output formats remains a hurdle, our work underscores ChatGPT's proficiency in deriving insightful inferences from the text. Our findings concerning the influence of few-shot learning examples on output bias also suggest the need for

deeper investigation into the interplay between example selection and model performance.

With reliable automation methods that can identify when a person is describing a personal experience with LKD, future work will extend the reach to additional media platforms, each of which has its own system for reaching users via advertising. There will certainly be potential biases in accessing educational information about living donations based on the characteristics of audiences most likely to post on each platform. To not exacerbate disparities, one must examine the generalizability of the profiles across multiple platforms and ensure the dissemination of information across platforms that reach diverse audiences and non-English speakers. An examination of access to most audience members, particularly the underserved, is warranted to ensure that all communities are reached equitably.

Utility of Results

By identifying these unique user classifications, tailored educational interventions for different profiles could be designed. First, for those most actively considering living donation, there could be social media campaigns built and targeted to specific users to invite them to learn more about living donation. These users can be referred to a trusted site, which includes education materials and an opportunity to register to begin donor medical evaluation at a nearby transplant center [41,42]. For individuals discussing their concerns about the costs involved with becoming a living donor, referrals to websites that discuss the ways to apply for grants to cover the out-of-pocket costs and lost wages could be valuable in their decision-making [70].

Second, for donors and families identified to have completed donations, campaigns inviting them to share their experiences on a living donor storytelling website [8,9] might result in more real-life stories being captured from diverse individuals to increase awareness of living donations for the national public. Stories are particularly valuable for educating learners with low health literacy or those for whom English is not their primary language about the possibilities of living donation [71].

Finally, it will be very important to work with experts in marketing and campaign design to plan social media campaigns that are motivating and helpful for patients and their families at different points along their donation journey. Identifying motivated learners from platforms such as Reddit, delivering content to them about living donation, and assessing its impact on learning more or pursuing donation are our next planned steps.

The proposed profiles may incorrectly identify a person's interest or stage of pursuit of donation, making any educational information sent to them irrelevant. In contrast, users could also be made uncomfortable if the education being provided matches their needs perfectly, indicating that their data are being scrutinized. Users can always disregard nonrelevant content; however, it will be important in the design of new campaigns not to assume with too much certainty that all learners are correctly identified. Respect for users is an ethical tenet that must always be considered in designing the campaigns and communicating how we found that they might be considering living donations as we move forward.

Conclusions

Much of the previous health care–related research about LLMs has been centered on their reliability in producing quality medical information. In contrast, we endeavor to extract individual-level information from the internet that can be used to inform health care providers. Consequently, there is little comparison that can be made to previous work other than to say that the reliability of the models is subject to the instructions they are given. However, our experimental results do illustrate that when using examples as part of the prompt (few-shot), bias toward the class of the given examples can affect performance. We have also shown that simulating a dialogue between 2 experts is more effective than using stand-alone reasoning.

This study takes a significant step in applying advanced NLP methods to the field of LKD, focusing on automating the detection of personal LKD experiences in online content. Both BERT and ChatGPT proved effective for this task, each with its own advantages and disadvantages. Our new DUCC method outperformed traditional reasoning approaches, emphasizing the importance of further work on improving prompt design. The study also highlights the need for automated prompt creation to reduce the time and effort currently required for manual testing, making NLP applications in the LKD field more efficient and impactful.

Acknowledgments

This study is supported in part by the Logistics and Distribution Institute at the University of Louisville. XC is supported by the American Heart Association (23CSA1052735), and National Science Foundation (CMMI-2430998).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full experimental results.
[XLSX File (Microsoft Excel File), 13 KB - ai_v4i1e57319_app1.xlsx]

References

1. Abecassis M, Bartlett ST, Collins AJ, Davis CL, Delmonico FL, Friedewald JJ, et al. Kidney transplantation as primary therapy for end-stage renal disease: a National Kidney Foundation/Kidney Disease Outcomes Quality Initiative (NKF/KDOQITM) conference. Clin J Am Soc Nephrol 2008 Mar;3(2):471-480 [FREE Full text] [doi: 10.2215/CJN.05021107] [Medline: 18256371]
2. Axelrod DA, Schnitzler MA, Xiao H, Irish W, Tuttle-Newhall E, Chang S, et al. An economic assessment of contemporary kidney transplant practice. Am J Transplant 2018 May;18(5):1168-1176 [FREE Full text] [doi: 10.1111/ajt.14702] [Medline: 29451350]
3. All-time records again set in 2021 for organ transplants, organ donation from deceased donors. Health Resources and Services Administration. URL: https://optn.transplant.hrsa.gov/news/all-time-records-again-set-in-2021-for-organ-transplants-organ-donation-from-deceased-donors/ [accessed 2023-01-25]
4. Lentine KL, Smith JM, Hart A, Miller J, Skeans MA, Larkin L, et al. OPTN/SRTR 2020 annual data report: kidney. Am J Transplant 2022 Mar;22 Suppl 2:21-136 [FREE Full text] [doi: 10.1111/ajt.16982] [Medline: 35266618]
5. Purnell TS, Hall YN, Boulware LE. Understanding and overcoming barriers to living kidney donation among racial and ethnic minorities in the United States. Adv Chronic Kidney Dis 2012 Jul;19(4):244-251 [FREE Full text] [doi: 10.1053/j.ackd.2012.01.008] [Medline: 22732044]
6. Purnell TS, Luo X, Cooper LA, Massie AB, Kucirka LM, Henderson ML, et al. Association of race and ethnicity with live donor kidney transplantation in the United States from 1995 to 2014. JAMA 2018 Jan 02;319(1):49-61 [FREE Full text] [doi: 10.1001/jama.2017.19152] [Medline: 29297077]
7. Morgan SE, Harrison TR, Long SD, Afifi WA, Stephenson MS, Reichert T. Family discussions about organ donation: how the media influences opinions about donation decisions. Clin Transplant 2005 Oct 11;19(5):674-682. [doi: 10.1111/j.1399-0012.2005.00407.x] [Medline: 16146561]
8. Ho EW, Murillo AL, Davis LA, Iraheta YA, Advani SM, Feinsinger A, et al. Findings of living donation experiences shared on a digital storytelling platform: a thematic analysis. PEC Innov 2022 Dec;1:100023 [FREE Full text] [doi: 10.1016/j.pecinn.2022.100023] [Medline: 37213721]
9. Davis L, Iraheta YA, Ho EW, Murillo AL, Feinsinger A, Waterman AD. Living kidney donation stories and advice shared through a digital storytelling library: a qualitative thematic analysis. Kidney Med 2022 Jul;4(7):100486 [FREE Full text] [doi: 10.1016/j.xkme.2022.100486] [Medline: 35755303]
10. Kaplow K, Ruck JM, Levan ML, Thomas AG, Stewart D, Massie AB, et al. National attitudes towards living kidney donation in the United States: results of a public opinion survey. Kidney Med 2024 Mar;6(3):100788 [FREE Full text] [doi: 10.1016/j.xkme.2023.100788] [Medline: 38435064]
11. Amaral S, McCulloch CE, Black E, Winnicki E, Lee B, Roll GR, et al. Trends in living donation by race and ethnicity among children with end-stage renal disease in the United States, 1995-2015. Transplant Direct 2020 Jul;6(7):e570 [FREE Full text] [doi: 10.1097/TXD.0000000000001008] [Medline: 32766425]
12. Waterman AD, Morgievich M, Cohen DJ, Butt Z, Chakkera HA, Lindower C, American Society of Transplantation. Living donor kidney transplantation: improving education outside of transplant centers about live donor transplantation--recommendations from a consensus conference. Clin J Am Soc Nephrol 2015 Sep 04;10(9):1659-1669 [FREE Full text] [doi: 10.2215/CJN.00950115] [Medline: 26116651]
13. Waterman AD, Peipert JD. An explore transplant group randomized controlled education trial to increase dialysis patients' decision-making and pursuit of transplantation. Prog Transplant 2018 Jun 26;28(2):174-183. [doi: 10.1177/1526924818765815] [Medline: 29699451]
14. Asghari M, Nielsen J, Gentili M, Koizumi N, Elmaghraby A. Classifying comments on social media related to living kidney donation: machine learning training and validation study. JMIR Med Inform 2022 Nov 08;10(11):e37884 [FREE Full text] [doi: 10.2196/37884] [Medline: 36346661]
15. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS'17; December 4-9, 2017; Long Beach, CA p. 6000-6010 URL: https://dl.acm.org/doi/10.5555/3295222.3295349
16. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: generalized autoregressive pretraining for language understanding. arXiv Preprint posted online June 19, 2019 [FREE Full text]
17. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv Preprint posted online July 26, 2019 [FREE Full text]
18. Chernyavskiy A, Ilvovsky D, Nakov P. Transformers: "the end of history" for NLP? arXiv Preprint posted online April 9, 2021 [FREE Full text]
19. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online October 11, 2018 [FREE Full text]
20. Acheampong FA, Nunoo-Mensah H, Chen W. Transformer models for text-based emotion detection: a review of BERT-based approaches. Artif Intell Rev 2021 Feb 08;54(8):5789-5829. [doi: 10.1007/S10462-021-09958-2]
21. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber Phys Syst 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]

22. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv Preprint posted online February 21, 2023 [FREE Full text]
23. Reynolds L, McDonell K. Prompt programming for large language models: beyond the few-shot paradigm. arXiv Preprint posted online February 15, 2021 [FREE Full text] [doi: 10.1145/3411763.3451760]
24. Liu Y, Deng G, Xu Z, Li Y, Zheng Y, Zhang Y, et al. Jailbreaking ChatGPT via prompt engineering: an empirical study. arXiv Preprint posted online May 23, 2023 [FREE Full text]
25. Shi Z, Wang Y, Yin F, Chen X, Chang KW, Hsieh CJ. Red teaming language model detectors with language models. arXiv Preprint posted online May 31, 2023 [FREE Full text] [doi: 10.1162/tacl_a_00639]
26. Casper S, Lin J, Kwon J, Cilp G, Hadfield-Menell D. Explore, establish, exploit: red teaming language models from scratch. arXiv Preprint posted online June 15, 2023 [FREE Full text]
27. Shinn N, Cassano F, Berman E, Gopinath A, Narasimhan K, Yao S. Reflexion: language agents with verbal reinforcement learning. arXiv Preprint posted online March 20, 2023 [FREE Full text]
28. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv Preprint posted online January 28, 2022 [FREE Full text]
29. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-consistency improves chain of thought reasoning in language models. arXiv Preprint posted online March 21, 2021 [FREE Full text]
30. Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, et al. Tree of thoughts: deliberate problem solving with large language models. arXiv Preprint posted online May 17, 2023 [FREE Full text]
31. Papers. Prompt Engineering Guide. URL: https://www.promptingguide.ai/papers [accessed 2024-04-29]
32. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 04;25:e50638 [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
33. Wang J, Shi E, Yu S, Wu Z, Ma C, Dai H, et al. Prompt engineering for healthcare: methodologies and applications. arXiv Preprint posted online April 28, 2023 [FREE Full text]
34. Wang L, Chen X, Deng X, Wen H, You M, Liu W, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med 2024 Feb 20;7(1):41 [FREE Full text] [doi: 10.1038/s41746-024-01029-4] [Medline: 38378899]
35. Patel D, Raut G, Zimlichman E, Cheetirala SN, Nadkarni G, Glicksberg BS, et al. The limits of prompt engineering in medical problem-solving: a comparative analysis with ChatGPT on calculation based USMLE medical questions. medRxiv Preprint posted online August 9, 2023 [FREE Full text] [doi: 10.1101/2023.08.06.23293710]
36. Lim S, Schmälzle R. Artificial intelligence for health message generation: an empirical study using a large language model (LLM) and prompt engineering. Front Commun 2023 May 26;8:1129082. [doi: 10.3389/fcomm.2023.1129082]
37. Ali H, Bulbul MF, Shah Z. Prompt engineering in medical image segmentation: an overview of the paradigm shift. In: Proceedings of the 2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things. 2023 Presented at: AIBThings '23; September 16-17, 2023; Mount Pleasant, MI p. 1-4 URL: https://ieeexplore.ieee.org/document/10292475 [doi: 10.1109/aibthings58340.2023.10292475]
38. Argyris YA, Monu K, Tan P, Aarts C, Jiang F, Wiseley KA. Using machine learning to compare provaccine and antivaccine discourse among the public on social media: algorithm development study. JMIR Public Health Surveill 2021 Jun 24;7(6):e23105 [FREE Full text] [doi: 10.2196/23105] [Medline: 34185004]
39. Henderson ML. Social media in the identification of living kidney donors: platforms, tools, and strategies. Curr Transpl Rep 2018 Jan 18;5(1):19-26. [doi: 10.1007/S40472-018-0179-8]
40. Jiang X, Jiang W, Cai J, Su Q, Zhou Z, He L, et al. Characterizing media content and effects of organ donation on a social media platform: content analysis. J Med Internet Res 2019 Mar 12;21(3):e13058 [FREE Full text] [doi: 10.2196/13058] [Medline: 30860489]
41. DuBray BJ, Shawar SH, Rega SA, Smith KM, Centanni KM, Warmke K, et al. Impact of social media on self-referral patterns for living kidney donation. Kidney360 2020 Dec 31;1(12):1419-1425. [doi: 10.34067/kid.0003212020]
42. Joachim E. Self-referral patterns of living kidney donors via social media: examining an expanding platform. Kidney360 2020 Dec 31;1(12):1337-1338 [FREE Full text] [doi: 10.34067/KID.0005732020] [Medline: 35372901]
43. Kumar K, King E, Muzaale A, Konel J, Bramstedt K, Massie A, et al. A smartphone app for increasing live organ donation. Am J Transplant 2016 Dec;16(12):3548-3553 [FREE Full text] [doi: 10.1111/ajt.13961] [Medline: 27402293]
44. Murphy MD, Pinheiro D, Iyengar R, Lim G, Menezes R, Cadeiras M. A data-driven social network intervention for improving organ donation awareness among minorities: analysis and optimization of a cross-sectional study. J Med Internet Res 2020 Jan 14;22(1):e14605 [FREE Full text] [doi: 10.2196/14605] [Medline: 31934867]
45. Kazley AS, Hamidi B, Balliet W, Baliga P. Social media use among living kidney donors and recipients: survey on current practice and potential. J Med Internet Res 2016 Dec 20;18(12):e328 [FREE Full text] [doi: 10.2196/jmir.6176] [Medline: 27998880]
46. Ruck JM, Henderson ML, Eno AK, Van Pilsum Rasmussen SE, DiBrito SR, Thomas AG, et al. Use of Twitter in communicating living solid organ donation information to the public: an exploratory study of living donors and transplant professionals. Clin Transplant 2019 Jan 07;33(1):e13447 [FREE Full text] [doi: 10.1111/ctr.13447] [Medline: 30421841]

47. Eno AK, Thomas AG, Ruck JM, Van Pilsum Rasmussen SE, Halpern SE, Waldram MM, et al. Assessing the attitudes and perceptions regarding the use of mobile health technologies for living kidney donor follow-up: survey study. JMIR Mhealth Uhealth 2018 Oct 09;6(10):e11192 [FREE Full text] [doi: 10.2196/11192] [Medline: 30305260]
48. Gordon EJ, Shand J, Black A. Google analytics of a pilot mass and social media campaign targeting Hispanics about living kidney donation. Internet Interv 2016 Nov;6:40-49 [FREE Full text] [doi: 10.1016/j.invent.2016.09.002] [Medline: 30135813]
49. Britt RK, Britt BC, Anderson J, Fahrenwald N, Harming S. "Sharing hope and healing": a culturally tailored social media campaign to promote living kidney donation and transplantation among native Americans. Health Promot Pract 2021 Nov 02;22(6):786-795. [doi: 10.1177/1524839920974580] [Medline: 33267677]
50. Pacheco DF, Pinheiro D, Cadeiras M, Menezes R. Characterizing organ donation awareness from social media. In: Proceedings of the 33rd International Conference on Data Engineering. 2017 Presented at: ICDE '17; April 19-22, 2017; San Diego, CA p. 1541-1548 URL: https://ieeexplore.ieee.org/document/7930122 [doi: 10.1109/icde.2017.225]
51. Basu G, Nair S, Sibel G, Dheerendra P, Penmatsa KR, Balasubramanian K, et al. Social media and organ donation - a narrative review. Indian J Transplant 2021;15(2):139-146 [FREE Full text] [doi: 10.4103/ijot.ijot_138_20]
52. Tan M, Mulloy M, Pollinger H, Gibney E. Impact of social media on living kidney donation awareness. Transplantation 2014;98:836-837. [doi: 10.1097/00007890-201407151-02857]
53. Chang A, Anderson EE, Turner HT, Shoham D, Hou SH, Grams M. Identifying potential kidney donors using social networking web sites. Clin Transplant 2013 Apr 22;27(3):E320-E326 [FREE Full text] [doi: 10.1111/ctr.12122] [Medline: 23600791]
54. Ayorinde JO, Saeb-Parsy K, Hossain A. Opportunities and challenges in using social media in organ donation. JAMA Surg 2020 Sep 01;155(9):797-798. [doi: 10.1001/jamasurg.2020.0791] [Medline: 32936283]
55. Lee C, Lin M, Lin H, Ting Y, Wang H, Wang C, et al. Survey of factors associated with the willingness toward living kidney donation. J Formos Med Assoc 2022 Nov;121(11):2300-2307 [FREE Full text] [doi: 10.1016/j.jfma.2022.06.007] [Medline: 35803885]
56. Garcia Valencia OA, Thongprayoon C, Miao J, Suppadungsuk S, Krisanapan P, Craici IM, et al. Empowering inclusivity: improving readability of living kidney donation information with ChatGPT. Front Digit Health 2024 Apr 10;6:1366967 [FREE Full text] [doi: 10.3389/fdgth.2024.1366967] [Medline: 38659656]
57. Wu X, Zheng W, Xia X, Lo D. Data quality matters: a case study on data label correctness for security bug report prediction. IIEEE Trans Software Eng 2022 Jul 1;48(7):2541-2556. [doi: 10.1109/tse.2021.3063727]
58. Lin TY, Goyel P, Girshick R, He K, Dollár P. Focal loss for dense object detection. arXiv Preprint posted online August 7, 2017 [FREE Full text] [doi: 10.1109/iccv.2017.324]
59. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv Preprint posted online November 14, 2017 [FREE Full text] [doi: 10.1090/mbk/121/79]
60. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019 Presented at: KDD '19; August 4-8, 2019; Anchorage, AK p. 2623-2631 URL: https://dl.acm.org/doi/10.1145/3292500.3330701 [doi: 10.1145/3292500.3330701]
61. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv Preprint posted online May 28, 2020 [FREE Full text]
62. OpenAI developer platform. OpenAI. URL: https://platform.openai.com [accessed 2024-04-29]
63. Zhou Y, Muresanu AI, Han Z, Paster K, Pitis S, Chan H, et al. Large language models are human-level prompt engineers. arXiv Preprint posted online November 3, 2022 [FREE Full text]
64. Pryzant R, Iter D, Li J, Lee YT, Zhu C, Zeng M. Automatic prompt optimization with "gradient descent" and beam search. arXiv Preprint posted online May 4, 2023 [FREE Full text] [doi: 10.18653/v1/2023.emnlp-main.494]
65. Sordoni A, Yuan X, Côté MA, Pereira M, Trischler A, Xiao Z, et al. Joint prompt optimization of stacked LLMs using variational inference. arXiv Preprint posted online June 21, 2023 [FREE Full text]
66. Sun H, Li X, Xu Y, Homma Y, Cao Q, Wu M, et al. AutoHint: automatic prompt optimization with hint generation. arXiv Preprint posted online July 13, 2023 [FREE Full text]
67. Yang C, Wang X, Lu Y, Liu H, Le QV, Zhou D, et al. Large language models as optimizers. arXiv Preprint posted online September 7, 2023 [FREE Full text]
68. Chen A, Dohan DM, So DR. EvoPrompting: language models for code-level neural architecture search. arXiv Preprint posted online February 28, 2023 [FREE Full text]
69. Fernando C, Banarse H, Michalewski H, Osindero S, Rocktäschel T. Promptbreeder: self-referential self-improvement via prompt evolution. arXiv Preprint posted online September 28, 2023 [FREE Full text]
70. Home. National Living Donor Assistance Center. URL: https://www.livingdonorassistance.org/ [accessed 2025-09-01]
71. Lipsey AF, Waterman AD, Wood EH, Balliet W. Evaluation of first-person storytelling on changing health-related attitudes, knowledge, behaviors, and outcomes: a scoping review. Patient Educ Couns 2020 Oct;103(10):1922-1934. [doi: 10.1016/j.pec.2020.04.014] [Medline: 32359877]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
DUCC: dialogue until classification consensus
LDKT: living donor kidney transplantation
LKD: living kidney donation
LLM: large language model
NLP: natural language processing

Edited by S Gardezi, F Dankar; submitted 12.02.24; peer-reviewed by GK Gupta, A Hassan, W Cheungpasitporn; comments to author 28.08.24; revised version received 18.09.24; accepted 18.11.24; published 07.02.25.

Please cite as:
Nielsen J, Chen X, Davis L, Waterman A, Gentili M
Investigating the Classification of Living Kidney Donation Experiences on Reddit and Understanding the Sensitivity of ChatGPT to Prompt Engineering: Content Analysis
JMIR AI 2025;4:e57319
URL: https://ai.jmir.org/2025/1/e57319
doi:10.2196/57319
PMID:39918869

Research Letter

Evaluating ChatGPT's Efficacy in Pediatric Pneumonia Detection From Chest X-Rays: Comparative Analysis of Specialized AI Models

Nitin Chetla[1], BS; Mihir Tandon[2], BA; Joseph Chang[3], BS; Kunal Sukhija[4], MD; Romil Patel[1], BS; Ramon Sanchez[5], MD

[1]Department of Radiology, University of Virginia School of Medicine, Charlottesville, VA, United States

[2]Department of Orthopaedics, Albany Medical College, Albany, NY, United States

[3]Department of Radiology, University of Passau, Passau, Germany

[4]Department of Emergency Medicine, Kaweah Health Medical Center, Visalia, CA, United States

[5]Department of Radiology, Children's National Hospital, Washington, DC, United States

Corresponding Author:
Mihir Tandon, BA
Department of Orthopaedics
Albany Medical College
43 New Scotland Ave
Albany, NY, 12208
United States
Phone: 1 3322488708
Email: tandonm@amc.edu

(JMIR AI 2025;4:e67621) doi:10.2196/67621

KEYWORDS

artificial intelligence; ChatGPT; pneumonia; chest x-ray; pediatric; radiology; large language models; machine learning; pneumonia detection; diagnosis; pediatric pneumonia

Introduction

Recent studies have demonstrated the versatility of ChatGPT in health care [1]. In contrast, convolutional neural networks (CNNs) have an established history in medical imaging, particularly in identifying pneumonia from chest x-rays. CNNs are a class of deep learning algorithms that recognize patterns in images, making them invaluable tools in radiology and other imaging-based diagnostics [2]. Numerous studies demonstrate CNNs' effectiveness in medical imaging [3].

With advancements and developments in artificial intelligence (AI) technology, this research aims to evaluate the effectiveness of using ChatGPT-4 to detect pneumonia on x-ray images and compare its performance with specialized CNNs. These technologies could address radiologist shortages.

Community-acquired pneumonia incidence has reached 450 million cases worldwide annually [4]. In diagnosing pneumonia, a clinical history, physical examination, and laboratory tests are required, but clinical guidelines consider chest x-ray as the gold standard for distinguishing pneumonia from other respiratory tract infections [5]. However, interobserver agreement has been poor in chest radiographs of pediatric pneumonia [6]. Technological improvements such as ChatGPT and AI can help detect and diagnose pediatric pneumonia.

Methods

This study used a dataset of chest x-rays from the Kaggle dataset "Chest X-Ray Images (Pneumonia)," originally sourced from the Guangzhou Women and Children's Medical Center [3,7]. The dataset consists of 5863 pneumonia and normal chest x-ray images. The images were selected from retrospective cohorts of pediatric patients, aged 1-5 years, who underwent anterior-posterior chest x-rays as part of their workup. For quality assurance, the diagnoses associated with the images were graded by three expert physicians. The dataset includes bacterial and viral pneumonia cases but does not specify the type of pneumonia or distinguish between simple and complicated pneumonia.

The study used a subset of this dataset, consisting of 500 x-rays with pneumonia and 500 without pneumonia. Each image is stored in a subfolder labeled "Pneumonia" or "Normal," enabling straightforward categorization and access. ChatGPT-4 was then prompted with "Based on the image, does the patient have A) pneumonia or B) no pneumonia? Only output the answer as A or B." The results were analyzed.

Results

ChatGPT-4 Turbo was biased toward the answer nonpneumonia (Table 1 and Figure 1). The substantial bias affects the statistical measures used. ChatGPT-4o performs slightly better overall, except in sensitivity and specificity.

Figure 1. Confusion matrix of ChatGPT-4 Turbo.

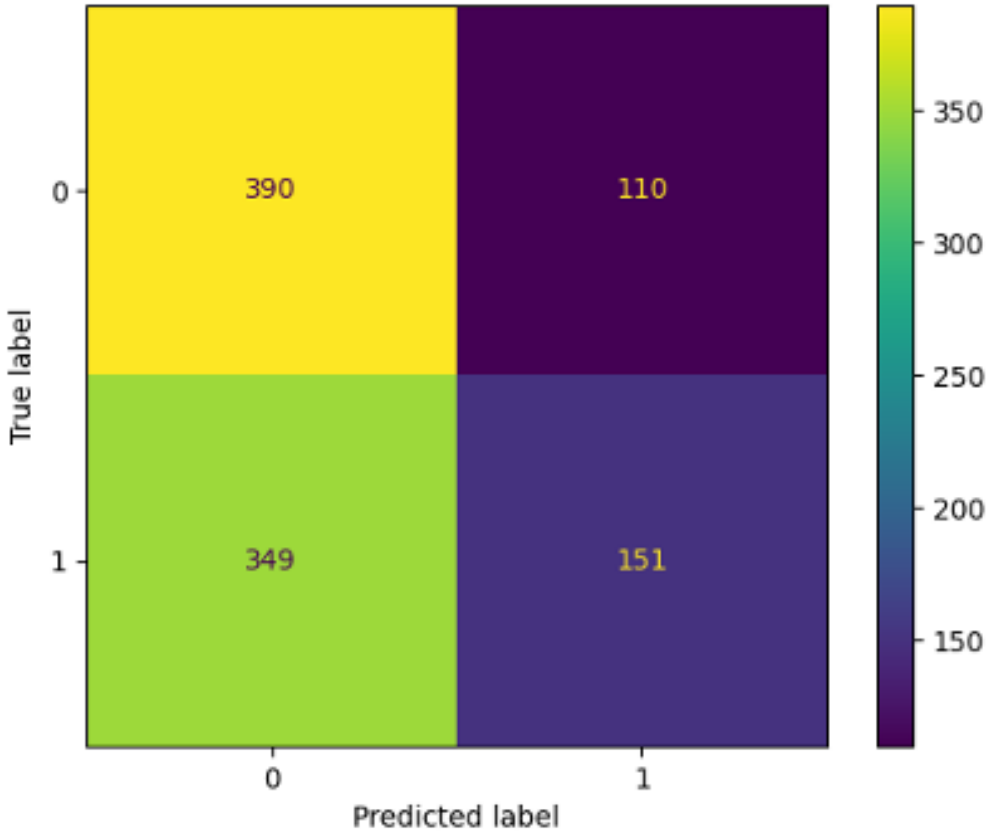


Table 1. Statistical overview table of results of ChatGPT-4 Turbo and GPT-4o.

Statistic	ChatGPT-4 Turbo	ChatGPT-4o
Accuracy (95% CI)	0.541 (0.511-0.571)	0.612 (0.582-0.642)
Precision (95% CI)	0.579 (0.548-0.607)	0.576 (0.545-0.607)
Specificity (95% CI)	0.780 (0.754-0.806)	0.839 (0.816-0.861)
Sensitivity (95% CI)	0.302 (0.274-0.333)	0.850 (0.828-0.872)
F_1-score (95% CI)	0.397 (0.367-0.427)	0.685 (0.656-0.714)

Discussion

Although ChatGPT-4 Turbo demonstrated a slight ability to differentiate between pneumonia and nonpneumonia cases, this accuracy was overshadowed by the model's strong bias, making its distinction between the two classes unreliable for clinical use. ChatGPT-4o is equally unreliable for clinical use.

Compared with Kermany et al [3], our ChatGPT results are subpar. ChatGPT's best accuracy was 61.2% (ChatGPT-4o) in this study, compared to 92.8%. ChatGPT-4o's sensitivity and specificity were also lower in this study: 85% and 38% compared to 93.2% and 90.1%, respectively. Noticeably, ChatGPT-4o's specificity was very low comparatively. ChatGPT-4 Turbo's sensitivity and specificity results were nearly reversed compared to its successor, indicating a substantial shift in predictive behavior. Our experiment only involved 1000 testing samples in total, while Kermany et al [3] trained with 5232 samples and tested another 624 samples.

Several challenges exist in using ChatGPT-4 Turbo for diagnosing pneumonia from chest x-ray radiographs. The model's strong bias toward classifying images as nonpneumonia significantly affected the accuracy and other measures used to evaluate the model's performance. The high number of false negatives could lead to delayed or missed diagnoses in a clinical setting.

A limitation of this study is that the lack of complex pattern recognition of pediatric pneumonia by ChatGPT may be anticipated as the program has likely not been fine-tuned to assess these types of patterns. However, numerous studies have mentioned that programs like ChatGPT may replace radiologists,

but studies are needed to improve these programs, and radiologists will continue to be vital to health care [8]. By providing empirical evidence of the limitations of generalist AI models, this study underscores the need for task-specific fine-tuning and integration with computer vision models, which can help further develop these programs.

ChatGPT-4 has limitations when diagnosing pneumonia from chest x-ray radiographs as shown by this research. The model's strong bias toward a nonpneumonia diagnosis, limited ability to distinguish between the two classes, and lack of specialized medical knowledge suggest that it may be unsuitable for clinical use currently. Further research and development are needed to address these limitations and explore the potential of integrating language models with other computer vision techniques to improve the accuracy and reliability of automated pneumonia diagnosis from chest x-rays.

Conflicts of Interest

None declared.

References

1. Tan S, Xin X, Wu D. ChatGPT in medicine: prospects and challenges: a review article. Int J Surg 2024 Jun 01;110(6):3701-3706 [FREE Full text] [doi: 10.1097/JS9.0000000000001312] [Medline: 38502861]
2. Li M, Jiang Y, Zhang Y, Zhu H. Medical image analysis using deep learning algorithms. Front Public Health 2023;11:1273253 [FREE Full text] [doi: 10.3389/fpubh.2023.1273253] [Medline: 38026291]
3. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018 Feb 22;172(5):1122-1131.e9 [FREE Full text] [doi: 10.1016/j.cell.2018.02.010] [Medline: 29474911]
4. Sattar S, Nguyen A, Sharma S. Bacterial pneumonia. In: StatPearls. Treasure Island, FL: StatPearls Publishing; 2024.
5. Htun TP, Sun Y, Chua HL, Pang J. Clinical features for diagnosis of pneumonia among adults in primary care setting: a systematic and meta-review. Sci Rep 2019 May 20;9(1):7600. [doi: 10.1038/s41598-019-44145-y] [Medline: 31110214]
6. Voigt GM, Thiele D, Wetzke M, Weidemann J, Parpatt P, Welte T, et al. Interobserver agreement in interpretation of chest radiographs for pediatric community acquired pneumonia: findings of the pedCAPNETZ-cohort. Pediatr Pulmonol 2021 Aug;56(8):2676-2685 [FREE Full text] [doi: 10.1002/ppul.25528] [Medline: 34076967]
7. Mooney P. Chest x-ray images (pneumonia). Kaggle. URL: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia [accessed 2024-12-18]
8. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT. Diagn Interv Imaging 2023 Jun;104(6):269-274 [FREE Full text] [doi: 10.1016/j.diii.2023.02.003] [Medline: 36858933]

Abbreviations

AI: artificial intelligence
CNN: convolutional neural network

Edited by Y Huo; submitted 16.10.24; peer-reviewed by CH Chan; comments to author 23.11.24; revised version received 24.11.24; accepted 04.12.24; published 10.01.25.

Please cite as:
Chetla N, Tandon M, Chang J, Sukhija K, Patel R, Sanchez R
Evaluating ChatGPT's Efficacy in Pediatric Pneumonia Detection From Chest X-Rays: Comparative Analysis of Specialized AI Models
JMIR AI 2025;4:e67621
URL: https://ai.jmir.org/2025/1/e67621
doi:10.2196/67621
PMID:

Original Paper

Enhancing Interpretable, Transparent, and Unobtrusive Detection of Acute Marijuana Intoxication in Natural Environments: Harnessing Smart Devices and Explainable AI to Empower Just-In-Time Adaptive Interventions: Longitudinal Observational Study

Sang Won Bae[1], PhD; Tammy Chung[2], PhD; Tongze Zhang[1], MSc; Anind K Dey[3], PhD; Rahul Islam[1], BSc

[1]Human-Computer Interaction and Human-Centered AI Systems Lab, AI for Healthcare Lab, Charles V. Schaefer, Jr. School of Engineering and Science, Stevens Institute of Technology, Hoboken, NJ, United States

[2]Institute for Health, Healthcare Policy and Aging Research, Rutgers University, Newark, NJ, United States

[3]Information School, University of Washington, Seattle, WA, United States

Corresponding Author:
Sang Won Bae, PhD
Human-Computer Interaction and Human-Centered AI Systems Lab
AI for Healthcare Lab, Charles V. Schaefer, Jr. School of Engineering and Science
Stevens Institute of Technology
1 Castle Point Terrace
Hoboken, NJ, 07030-5906
United States
Phone: 1 4122658616
Email: sbae4@stevens.edu

Abstract

Background: Acute marijuana intoxication can impair motor skills and cognitive functions such as attention and information processing. However, traditional tests, like blood, urine, and saliva, fail to accurately detect acute marijuana intoxication in real time.

Objective: This study aims to explore whether integrating smartphone-based sensors with readily accessible wearable activity trackers, like Fitbit, can enhance the detection of acute marijuana intoxication in naturalistic settings. No previous research has investigated the effectiveness of passive sensing technologies for enhancing algorithm accuracy or enhancing the interpretability of digital phenotyping through explainable artificial intelligence in real-life scenarios. This approach aims to provide insights into how individuals interact with digital devices during algorithmic decision-making, particularly for detecting moderate to intensive marijuana intoxication in real-world contexts.

Methods: Sensor data from smartphones and Fitbits, along with self-reported marijuana use, were collected from 33 young adults over a 30-day period using the experience sampling method. Participants rated their level of intoxication on a scale from 1 to 10 within 15 minutes of consuming marijuana and during 3 daily semirandom prompts. The ratings were categorized as not intoxicated (0), low (1-3), and moderate to intense intoxication (4-10). The study analyzed the performance of models using mobile phone data only, Fitbit data only, and a combination of both (MobiFit) in detecting acute marijuana intoxication.

Results: The eXtreme Gradient Boosting Machine classifier showed that the MobiFit model, which combines mobile phone and wearable device data, achieved 99% accuracy (area under the curve=0.99; F_1-score=0.85) in detecting acute marijuana intoxication in natural environments. The F_1-score indicated significant improvements in sensitivity and specificity for the combined MobiFit model compared to using mobile or Fitbit data alone. Explainable artificial intelligence revealed that moderate to intense self-reported marijuana intoxication was associated with specific smartphone and Fitbit metrics, including elevated minimum heart rate, reduced macromovement, and increased noise energy around participants.

Conclusions: This study demonstrates the potential of using smartphone sensors and wearable devices for interpretable, transparent, and unobtrusive monitoring of acute marijuana intoxication in daily life. Advanced algorithmic decision-making provides valuable insight into behavioral, physiological, and environmental factors that could support timely interventions to

reduce marijuana-related harm. Future real-world applications of these algorithms should be evaluated in collaboration with clinical experts to enhance their practicality and effectiveness.

(JMIR AI 2025;4:e52270) doi:10.2196/52270

KEYWORDS

digital phenotyping; smart devices; intoxication; smartphone-based sensors; wearables; mHealth; marijuana; cannabis; data collection; passive sensing; Fitbit; machine learning; eXtreme Gradient Boosting Machine classifier; XGBoost; algorithmic decision-making process; explainable artificial intelligence; XAI; artificial intelligence; JITAI; decision support; just-in-time adaptive interventions; experience sampling

Introduction

Background

Acute effects of marijuana use impair motor skills and cognitive functions, such as attention and information processing [1-3], leading to adverse outcomes like poor academic and work performance, as well as an increased risk of motor vehicle crashes and fatal collisions [2,4]. Delta-9 tetrahydrocannabinol (THC), the principal psychoactive constituent of marijuana, binds to brain receptors, inducing a feeling of “euphoria” or being “high” [5]. Given the risks associated with THC-induced impairment, there is a critical need to detect episodes of marijuana intoxication in real time in the natural environment.

Several studies have explored the use of phone sensors or wearable devices to detect acute marijuana consumption. For example, a laboratory study with 10 participants used smartphone sensors (accelerometer, gyroscope) to detect acute marijuana use (3% or 7% THC vs placebo) and found that gait analysis with a support vector machine model achieved 92% accuracy (F_1-score=0.93) [6]. Another study (n=1) developed an electrochemical biosensor ring that detected salivary THC (minimum of 0.5 µM) and blood alcohol levels (minimum of 0.2 mM) within three minutes [7]. However, these studies were conducted in controlled environments, highlighting the need for research on using smartphone and wearable sensors to detect acute marijuana use in nonlaboratory, natural settings.

Detecting marijuana use in daily life could enable Just-In-Time interventions to reduce harm, such as avoiding driving while intoxicated [8]. However, challenges exist in detecting acute marijuana-related intoxication [9]. THC could be detected in an individual’s blood or urine for several days after consumption depending on factors such as recency, frequency, and chronicity of use [10]. Thus, a person who tests positive for THC might not be intoxicated or impaired at the time of testing [10]. Existing testing methods (eg, blood, urine, saliva, and breath) are not suitable for real-time detection, as THC can remain detectable in the body for days after consumption, which does not necessarily indicate current impairment [10].

To address these limitations, our recent study [11] used passive sensing via smartphones, coupled with self-reported intoxication, to detect marijuana use with 90% accuracy, using sensor-derived data from mobile phones alongside temporal variables, including time of day and day of week. Building on these findings [11], this study explores the use of wearable devices (eg, Fitbit) to enhance detection capabilities by incorporating physiological indicators, thereby improving the accuracy and immediacy of identifying marijuana effects in natural environments.

Wearable device–reported heart rate (HR) was examined as a potential physiological indicator of acute marijuana intoxication, based on laboratory studies, showing a dose-dependent increase in resting HR shortly after smoking or vaping marijuana [12-14]. Specifically, laboratory research reports that within 2-3 minutes of smoking marijuana, there is an acute increase (20%-60% dose-dependent) in resting HR [13], which might represent a “physiological signal” of the onset of a marijuana smoking episode. HR peaks 10-15 minutes after reaching maximum THC levels, followed by a rapid decline [12-14]. While tolerance to this effect may develop (eg, from a mean increase of 44.6 to 6.6 beats per minute (bpm) after 18-20 days of use) with chronic use, [12-14]. The acute HR increases have been validated in laboratory settings but have remained unexplored in real-world contexts. This study examines using off-the-shelf wearable devices, such as Fitbit, to detect acute HR increases as a physiological signal potentially correlated with self-reported marijuana intoxication.

Research Objectives and Contributions

While laboratory studies have established the link between HR changes and marijuana intoxication [12-14], its applicability in real-world scenarios is unexplored. To address this gap, we propose that combining wearable device data with smartphone sensors could improve algorithms for detecting marijuana intoxication in real-life settings. To enhance the interpretability of our algorithms and provide insights for just-in-time adaptive interventions, we incorporated explainable artificial intelligence (XAI) into our machine-learning pipeline. XAI helps clarify the role of digital biomarkers associated with self-reported marijuana intoxication in natural environments.

This study aims to determine whether data from smartphones (eg, accelerometer and GPS) and wearable devices (eg, Fitbit) can detect self-reported marijuana intoxication (“feeling high”) in the natural environment, a topic not previously investigated. Two hypotheses drive this research: (1) the novel MobiFit model, which combines smartphones and Fitbit data will outperform models that use only one data source in detecting self-reported intoxication; (2) HR and daily behavioral data (eg, step count) from Fitbit are important features for detecting self-reported marijuana intoxication. If either hypothesis is validated, it indicates the value of integrating wearable device data into daily life monitoring.

This study evaluates the performance of sensor-based models using (1) only smartphone sensors, (2) only Fitbit data, and (3)

the combined MobiFit model. We also used XAI to enhance understanding of key digital features from both smartphone sensors and Fitbit data associated with self-reported marijuana intoxication. Identifying smartphone-based sensors and Fitbit features that accurately detect self-reported marijuana intoxication in natural environments could ultimately trigger just-in-time interventions.

This study presents a comprehensive approach toward using mobile and wearable technology for detecting self-reported acute marijuana intoxication in real-life settings, emphasizing interpretability and transparency through XAI. This study demonstrates the potential of integrating smart devices with advanced analytical techniques to improve detection accuracy and support timely interventions based on detected intoxication levels.

Methods

Recruitment and Participants

A total of 57 participants aged 18-24 years were recruited through flyers, advertisements, and local communities. Eligibility criteria were (1) using marijuana at least twice a week, (2) owning a personal mobile phone, (3) not currently seeking treatment for substance abuse, (4) no self-reported history of psychosis, and (5) not taking any medication or using any medical device (eg, pacemaker) that could affect HR. Of the 57 participants, 24 participants were excluded from the analysis due to missing data (eg, no HR data and no mobile sensor data).

The final analysis focused on 33 participants aged 18-24 years, with an average age of 19.64 (SD 1.77) years. Among these, 23 participants identified as White, 4 participants as Black, and 6 participants as other race or ethnicity. The average age of first marijuana use was 16.48 (SD 1.84, range 13-22) years, and the average age of regular marijuana use was 17.03 (SD 1.72) years. In this subset, 24% (n=8) reported daily marijuana use, 9% (n=3) reported using it 5-6 times per week, and 67% (n=22) reported using it 2-4 times per week. Notably, 97% (n=32) of participants primarily used iOS smartphones, with only 3% (n=1) using Android devices.

Ethical Considerations

This naturalistic, observational follow-along study was approved by the university's institutional review board (Stevens 2020-008 [23-COAS3], Rutgers Pro2019002365). In line with similar Institutional Review Board–approved observational studies [15], all participants were informed about local medical and mental health resources. The study obtained a National Institutes of Health Certificate of Confidentiality. Written consent was obtained from participants, who were informed about privacy protections and the voluntary nature of their participation [16]. The research staff explained the types of data to be collected, the duration of data collection, and the purpose of the study.

Study Design

Participants completed a baseline laboratory assessment including interviews, questionnaires, and cognitive testing. They downloaded study apps from the App Store or Google Play Store to their smartphones. Research staff trained participants on how to use the apps and the study provided Fitbit Charge 2 for data collection. The AWARE mobile app [17] delivered experience sampling method (ESM) questions on marijuana use. Participants wore the Fitbit Charge 2 wristband to collect data on HR, physical activity (eg, step count), and sleep (eg, time, duration, and quality; see Table S2 in Multimedia Appendix 1 for Fitbit variables). The study collected continuous sensor data from smartphones and Fitbit devices, along with self-reported data on marijuana intoxication, for up to 30 days. A 30-day period was chosen to ensure sufficient data, given the study's inclusion criteria of frequent marijuana use. At the end of the study, participants completed a debriefing interview about their experience.

Participants were compensated for their time and effort, receiving US $75 for completing the baseline assessment, and US $25 for the debriefing interview. They earned US $10 for each day on which they completed more than 75% of data collection (eg, Fitbit and ESM).

Mobile Sensing Framework and Applications for Data Collection

AWARE App

AWARE is a mobile sensing framework [17] that passively and continuously collects data from smartphone sensors. This data can be used to infer human behavior patterns using various sensors: location (eg, distance traveled and circadian rhythm), physical movements (eg, acceleration and activity), device usage (eg, unlock, charge, keypress, and app usage), social patterns (eg, communication and conversations), and environmental context (eg, Wi-Fi, Bluetooth, sound or ambient noise, and light). The app, developed to track participants' natural behaviors in real-life settings, runs in the background 24/7 and collects sensor data with associated metadata, such as time stamps and communication logs. The data is transferred to a secure MySQL database owned and operated by the research team.

ESM

The mobile app also captured self-reports of marijuana use by participants. Two types of surveys were used [18]. Participants manually reported marijuana use within 15 minutes of consumption, detailing the amount used, mode of consumption, and the people whom the participant consumed marijuana with. They also rated their subjective intoxication on a scale from 0 (none) to 10 (a lot) [19]. Two hours later, the app prompted participants to complete an end-session survey indicating when intoxication symptoms subsided. In addition, fixed-time surveys were delivered daily at 10 AM, 3 PM, and 8 PM to collect information on the participants' daily lives, including time since last marijuana use, cravings, mood, and feelings (eg, relaxed, anxious, and sad), and other substance use (eg, alcohol and tobacco). Survey response windows were open for 5 hours to accommodate participants' schedules.

Fitbit Charge 2

Participants were provided with Fitbit Charge 2 devices and asked to wear them as much as possible. Fitbit collected

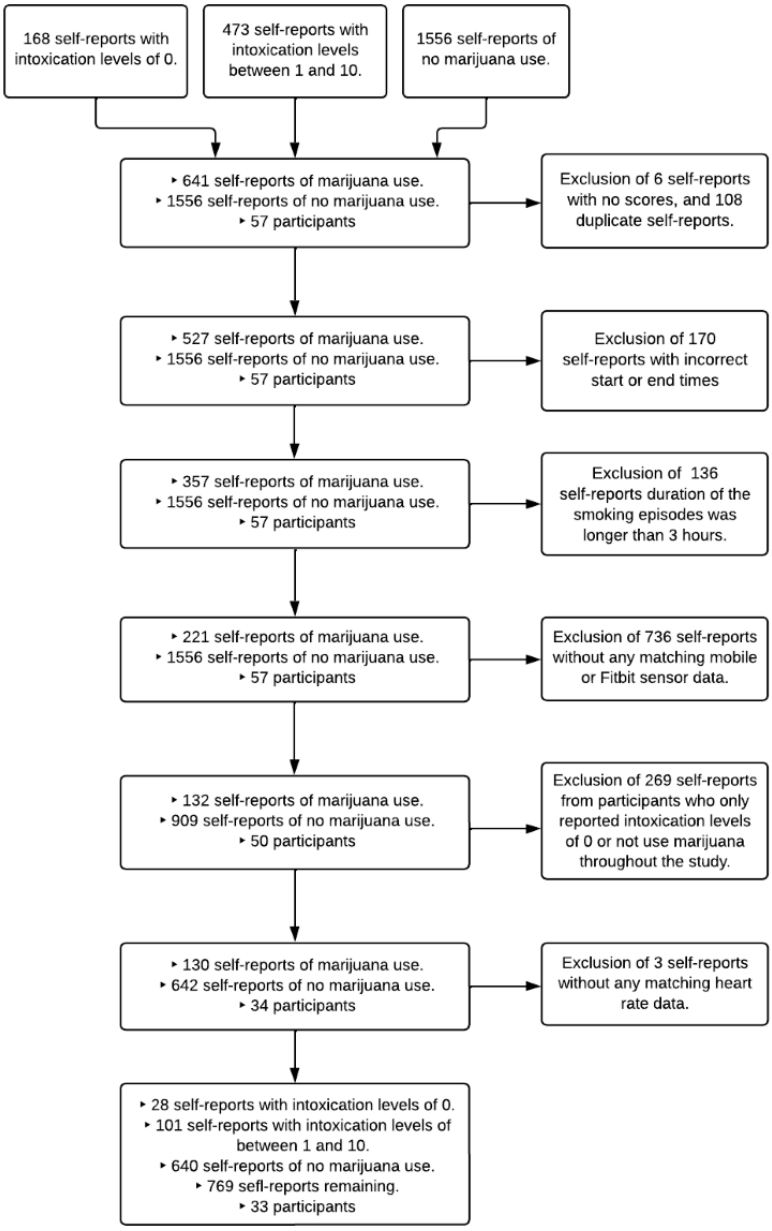
physiological data (eg, HR), activity data (eg, step count), and sleep. The study hypothesized that HR and behavioral data could signal episodes of acute marijuana intoxication. Fitbit data were retrieved from the Fitbit server at the end of the study using the Fitbit application programming interface.

Preparing Self-Report and Fitbit Data for Analysis

An episode of self-reported subjective marijuana intoxication was defined based on the ESM item: "How high are you feeling right now?" rated from 0 to 10 (0=not high to 10=a lot) [18,19]. To include episodes in the analysis, both start and end times had to be reported to calculate duration and label the sensor data. To capture behaviors without marijuana use, 1556 reports where participants answered "no" to the question "Did you smoke marijuana since the last report?" during afternoon (n=1151) and evening (n=950) surveys were labeled as "0" for the subjective rating of marijuana intoxication.

From all participants, we received 641 self-reports (mean 9.86, SD 8.49; median 7, IQR 4-13) and 1556 with no marijuana use reports (Figure 1). Out of 641 reports, 168 reports had a subjective intoxication rating of 0 and 10, and 6 reports had no rating. After excluding 6 reports without ratings and 108 duplicate reports, 527 samples remained. Reports with missing start and end times, or implausible episode durations (eg, longer than 3 hours) were excluded based on laboratory research indicating that smoked or vaped marijuana effects last less than 3 hours [20]. A total of 136 self-reports were excluded for exceeding this duration, leaving 1556 reports where no marijuana use was recorded [20].

Figure 1. Flowchart of participants and the data included in the analyses.



For model building, episodes without mobile sensor data (n=72) were excluded, leaving 221 marijuana self-reports. Furthermore, episodes without Fitbit sensor data (n=17) were excluded, leaving 50 participants. These participants provided 132

marijuana use self-reports and 909 “no marijuana use” reports. We analyzed reports from each participant, excluding those who only reported not using marijuana or had a rating of 0 for subjective intoxication, leaving a total of 642 with no marijuana use report or who reported 0 subjective intoxications when using marijuana and 34 people. Finally, to prevent participants from using Fitbit incorrectly, we excluded users without HR data, leaving a total of 33 people, who provided a total of 769 events: 640 “no marijuana use” reports and 129 marijuana use self-reports.

Extracting Smartphone and Fitbit Sensor Features

Following previous studies, we extracted audio features to detect social interactions [21,22] potentially associated with marijuana use. Audio features were extracted using the conversation plug-in, which detects whether a person was engaged in a conversation. Raw audio signals are converted to amplitude using the Euclidean norm [23], which categorizes ambient levels into silence, noise, voice, and unknown [24]. We also computed device use features, such as smartphone unlock minutes and the duration of device interaction sessions. In addition to audio features, we extracted GPS features to examine movement patterns related to marijuana use [25-28]. These included the radius of gyration, time at a location cluster, total distance traveled, number of clusters within a 5-minute window, acceleration, and phone angles. Environmental features, such as the number of Bluetooth devices detected, the most frequently contacted Wi-Fi access point, and light features (eg, average [avg], and maximum [max] lux) were also extracted. For most features, we calculated the minimum (min), max, avg, median (med), and SD. Further details on smartphone features can be found in Multimedia Appendix 1.

We used a 5-minute time window for extracting sensor feature statistics, as laboratory studies show a dose-dependent acute in resting HR within 2-3 minutes of marijuana use. Using larger time intervals could include data not related to marijuana use, given the average reported marijuana session duration is 75 (SD 46.2) minutes.

Raw data for HR, sleep, and steps were extracted from Fitbit. We first obtained per-minute HR and step count data using the Fitbit application programming interface. To exclude outliers, we refined data selection to omit instances where HR was below 40 bpm, as recommended by the American Heart Association [29,30]. We extracted feature statistics such as avg, SD, min, med, and max HR within a 5-minute window to explore the relationship between HR and marijuana intoxication levels (“moderate-intensive,” “low,” and “none”). Resting HR was defined as HR data collected when the participant was sedentary (ie, no steps taken) for more than 5 minutes. To further analyze HR patterns related to marijuana intoxication, we examined the degree of peakedness (kurtosis) and asymmetry (skewness) in HR data, as these features may reveal physiological changes associated with marijuana intoxication [31]. For more details, refer to Table S2 in Multimedia Appendix 2.

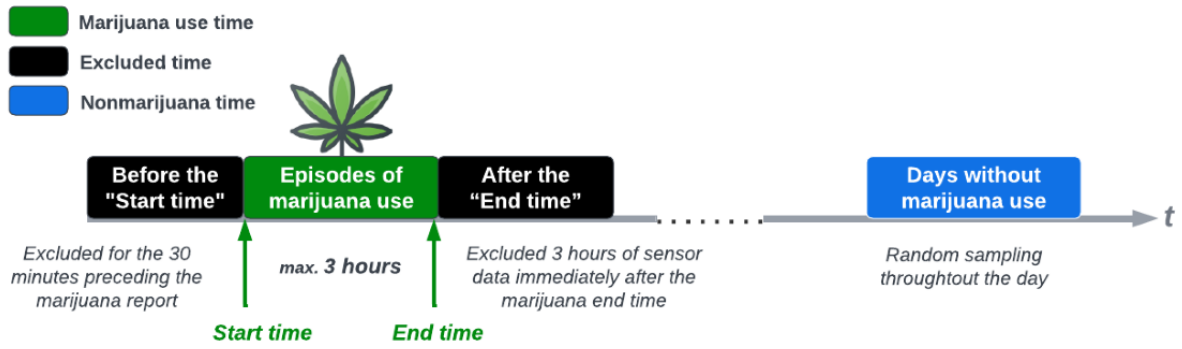
Ground Truth and Labeling Sensor Data

To accurately label the collected sensor data, we defined the duration of marijuana use episodes as those equal to or less than 3 hours, based on reported start and end times. We excluded 3 hours of sensor data following the reported end time to account for the continued effects of marijuana, even when participants reported a subjective intoxication level of 0. For example, if marijuana use was reported from 6 PM to 6:30 PM, data from 6:30 PM to 9:30 PM were excluded to account for residual effects. We also excluded data from 30 minutes before the reported start time to account for potential delays in self-reporting, based on pilot study findings that delays could range from 5 to 15 minutes. To collect nonmarijuana data, we randomly sampled sensor data from days when participants did not use marijuana (ie, nonmarijuana days). These samples were labeled using morning, afternoon, and evening surveys in which participants reported “no” to the ESM item “Did you smoke marijuana since the last report?” and indicated that the last use was more than 5 hours before the ESM time stamp (Figure 2).

We aimed to capture acute intoxication versus nonuse, classifying intoxication levels into three categories: 0 as “not intoxicated,” 1-3 as “low intoxication,” and 4-10 as “moderate-intensive intoxication” (MI). In total, we labeled 32,722 sensor stream samples (5-minute windows) as “not intoxicated” (154 from self-initiated survey coded as 0 high, and 32,586 from time-based self-reports), 423 samples as “low intoxication” (ratings between 1 and 3) and 772 samples as “moderate-intensive” (ratings between 4 and 10, with 10 indicating “a lot”).

Data from smartphones and Fitbit resulted in two datasets of different sizes. To ensure consistency, we down-sampled the smartphone dataset to include only samples overlapping with Fitbit data during the same time frames. This resulted in three datasets: (1) eXtreme Gradient Boosting (XGBoost)-Mobile: mobile phone only, (2) XGBoost-Fitbit: Fitbit-only, and (3) XGBoost-MobiFit: combined mobile and Fitbit data. The rationale for choosing Machine Learning (ML) models is detailed in Multimedia Appendix 3 and model comparison with different classifiers can be found in Multimedia Appendix 4.

Figure 2. Marijuana use episodes and labeling principle.

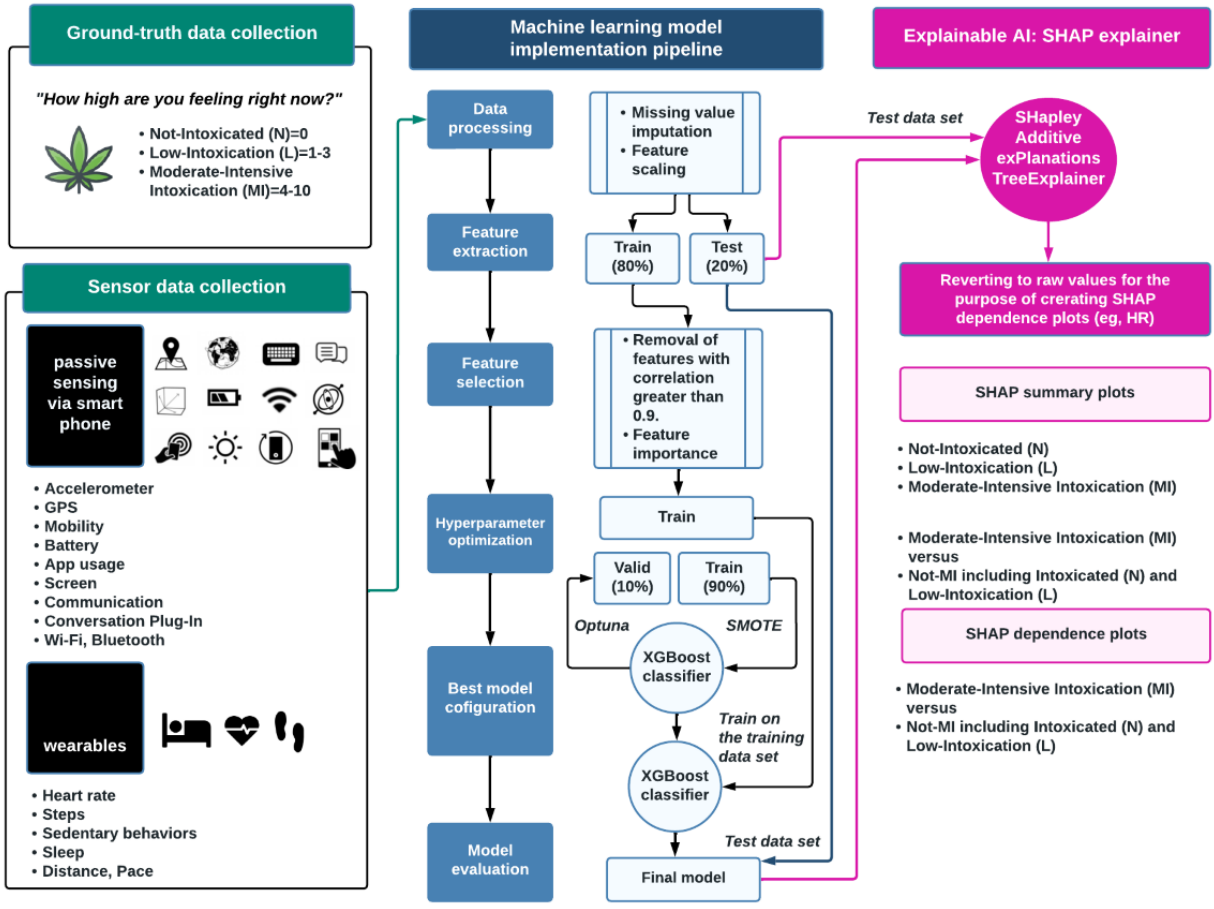


ML Pipeline

Feature Selection

We began data analysis by randomly partitioning the labeled sensor data into training (80%) and test (20% holdout) datasets. As shown in Figure 3, we first calculated Pearson correlation coefficients between features in the training dataset to identify highly covariant feature pairs (correlation coefficients >0.9) [32]. We then systematically removed one feature from each pair to reduce redundancy and improve model performance by retaining the most relevant and independent features. Next, we selected statistically significant features with a Gini coefficient importance [33] greater than 0.005. Details can be found in Multimedia Appendix 2.

Figure 3. Study overview. AI: artificial intelligence; HR: heart rate; SHAP: Shapley Additive exPlanations; SMOTE: Synthetic Minority Over-Sampling Technique; XGBoost: eXtreme Gradient Boosting Machine.



Hyper-Parameter Tuning and Cross-Validation

As shown in Figure 3, during hyper-parameter tuning in the training dataset, we used cross-validation to randomly leave 10% of the samples out, training the model on the remaining 90% and testing on the withheld 10%. We used the Synthetic Minority Over-Sampling Technique [34] to ensure equal representation across all classes. We further optimized model performance with a Bayesian-optimization-driven method called Optuna [35] to select the best combination of hyperparameters and 10-fold cross-validation on models with Optuna-optimized hyperparameters.

For the final model evaluation, we used the reserved test data (20% unseen data, as shown in Figure 3). The model was evaluated on predictions made on the test data. Finally, as shown in Figure 3 (right column), we conducted an XAI analysis to better understand the decision-making process of our final predictive model. We generated SHapley Additive exPlanations (SHAP) on the unseen test data to ensure our findings were explainable for data the model had not seen.

Model Evaluation Metrics

We evaluated model performance using F_1-score, recall, and precision, and selecting the best model based on the F_1-score [36]. Low precision indicates too many false positives (ie, detecting intoxication when there is none), here we would mistakenly intervene or notify the participant. Low recall indicates too many false negatives (ie, not detecting intoxication when it occurs), potentially leading to unsafe behaviors such as impaired driving. Therefore, while we prioritize the F_1-score, we also consider precision and recall.

Given our imbalanced samples, we used the area under the curve (AUC) metric, which provides a robust evaluation across all classification thresholds and is resilient to class imbalance.

XAI: Interpretation Approaches for Black-Box ML Models

To enhance algorithmic transparency, we used SHAP, a widely used interpretability method for ML models [37,38]. SHAP explains how specific data features influence model predictions, providing insights into the model's decision-making process. We identified the top 30 most significant features associated with marijuana intoxication reports, including their importance scores and visual summaries calculated by SHAP (see "Key Features Contributing to Model Performance" under the Results section). XGboost was selected due to its superior performance compared to other classifiers. The use of tree SHAP in this context reduces the computation time for SHAP values from exponential to polynomial [37].

Results

Timing, Duration, and Rating of Subjective Marijuana Intoxication

During the 30-day period, participants averaged 14 (SD 8.59) days of active participation. A total of 129 ESM self-initiated reports of marijuana use met the criteria for inclusion in the analysis: 101 reports of subjective marijuana intoxication (feeling high rated 1-10 out of 10) and 28 reports of feeling not high (0). Events not involving marijuana use were assigned a high rating of 0.

Tables 1 and 2 show the distribution of self-reported subjective marijuana intoxication across participants. Most episodes of intoxication (n=75) lasted between 30 minutes and 3 hours, with 54 episodes lasting up to 30 minutes (Table 1). Marijuana use was most often reported between 10 PM and 11 PM (n=24). Table 2 shows the distribution of ESM responses throughout the day. The average response latency to an ESM prompt expired. Most self-initiated reports of marijuana use occurred in the evenings: 14% (n=18) between 6 PM and 9 PM, and 39% (n=50) between 9 PM and midnight. On average, young adults rated their feeling of being high at 3.63 (SD 2.72) out of 10 when using marijuana (Table 3).

Table 1. Distribution of the duration of self-reported marijuana use episodes (n=129) across participants.

Duration[a] (hours)	Number of events
<0.5	54
<1	20
<1.5	23
<2.0	13
<2.5	13
<3	6

[a]Duration refers to the window of smoking episodes. From small (30 minutes) to relatively large windows (3 hours).

Table 2. Distribution of the start time of marijuana use episodes during the day (n=129).

Clock time (hours)	Number of events
0-1	7
1-2	8
2-3	2
3-4	0
4-5	0
5-6	0
6-7	0
7-8	1
8-9	0
9-10	5
10-11	8
11-12	2
12-13	6
13-14	6
14-15	5
15-16	4
16-17	3
17-18	4
18-19	5
19-20	6
20-21	7
21-22	10
22-23	24
23-0	16

Table 3. Distribution of self-reported “feeling high” during marijuana use.

High rating[a]	Number of events
0	28
1	9
2	9
3	17
4	14
5	14
6	17
7	10
8	7
9	4
10	0

[a]0-10 scale representing an intensity of feeling high, 10=a lot from the self-initiated reports of marijuana use. In our study, a value of 0 for the high report is labeled as “no-intoxication.”

Model Comparison: Mobile Only, Fitbit Only, and Mobile and Fitbit Integration

The first part of our analysis aimed to determine whether smartphone sensor features alone could be used for real-time detection of subjective marijuana intoxication and whether adding Fitbit data would improve model performance, justifying the added complexity of Fitbit data collection. We compared three ML models using the XGBoost classifier: (1) smartphone sensors only (XGBoost-Mobile), (2) Fitbit features only (XGBoost-Fitbit), and (3) a combined model using smartphone and Fitbit features (XGBoost-MobiFit).

Among the 3 models tested, the XGBoost-MobiFit model, which integrates smartphone and Fitbit data, had the best performance, achieving 99% accuracy, 92% precision, 79% recall, 85% F_1-score, and 99% AUC on the test dataset (Figure 4 and Table 4). These metrics indicate the XGBoost-MobiFit model's superior ability to accurately identify MI compared to low-intoxication and not-intoxicated states. While the XGBoost-Fitbit performed reasonably well, it did not match the performance of the XGBoost-MobiFit model in detecting marijuana intoxication. XGBoost-Fitbit achieved accuracy of 98%, 79% precision, 70% recall, 74% F_1-score, and 97% AUC. These results suggest that using only Fitbit data may not be as effective as combining it with smartphone sensor data for detecting subjective marijuana intoxication. Based on these findings, the added burden of wearing and charging the Fitbit device seems justified in future deployments. The combined model (XGBoost-MobiFit) demonstrated improved performance in detecting subjective marijuana intoxication compared to using smartphone or Fitbit data alone.

Combining Fitbit data with mobile data resulted in a significant improvement over the Fitbit-only model. The mobile-only model achieved an AUC of 96%, an F_1-score of 72%, a recall of 75%, and a precision of 70%. These results indicate that including Fitbit data adds value beyond what can be achieved with smartphone-based sensor data alone, as evidenced by a 13% improvement in F_1-score.

In summary, three key findings emerged: the XGBoost-Mobile model had the lowest performance (F_1-score=0.72, recall=0.75, precision=0.70); the XGBoost-Fitbit model (F_1-score=0.74, recall=0.70, precision=0.79) generally performed lower than the combined model; and the XGBoost-MobiFit model was the best performer with an F_1-score of 0.85, recall of 0.79, and precision of 0.92. As highlighted earlier, high precision and recall are critical so we focused on the F_1-score to identify the best-performing model. The model comparison with different classifiers is provided in Multimedia Appendix 4.

Figure 4. Model comparison to detect acute marijuana intoxication "low-intoxicated" (rating=1-3) versus "moderate-intensive intoxicated" (rating=4-10) versus "not-intoxicated" (rating=0). XGBoost-MobiFit: phone sensors and Fitbit (AUC=0.99; accuracy=0.99; left), XGBoost-Mobile: smartphone-based sensors (samples overlapping with Fitbit; AUC=0.96; accuracy=0.97; middle) and XGBoost-Fitbit: Fitbit only (AUC=0.97; accuracy=0.98; right). AUC: area under the curve; ROC: receiver-operating characteristic curve; XGBoost: eXtreme gradient boosting.

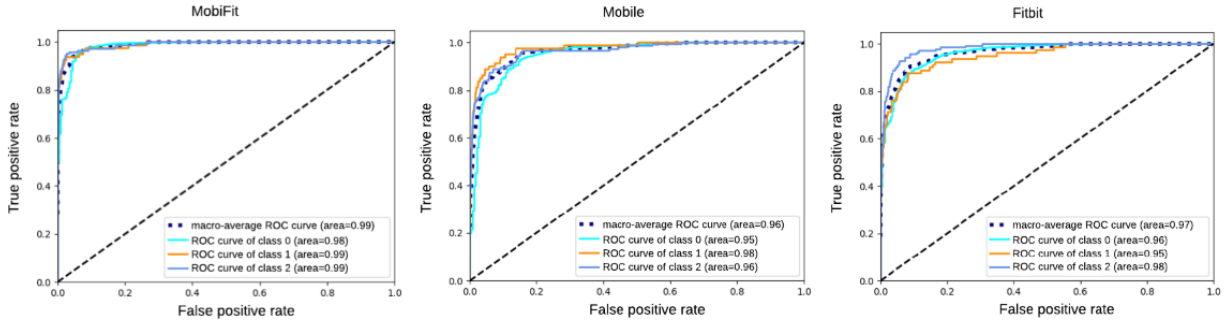


Table 4. Comparison of three XGBoost models using features selected in detecting moderate-intensive marijuana intoxication, low-intoxication, and not-intoxicated classes on the test dataset.

Machine learning model	AUC[a]	F_1-score	Recall	Precision	Accuracy
XGBoost-MobiFit	0.99	0.85	0.79	0.92	0.99
XGBoost-Mobile	0.96	0.72	0.75	0.70	0.97
XGBoost-Fitbit	0.97	0.74	0.70	0.79	0.98

[a]AUC: area under the curve.

Understanding Model Performance in Detecting the Risk State of "Moderate and Intensive Marijuana Intoxication"

For predicting the MI class alone, the MobiFit model outperformed the mobile and Fitbit-only models, exhibiting a substantial improvement in the F_1-score of 20% and 18%, respectively (Table 5). This improvement in F_1-score highlights the benefits of integrating data from both devices: enhanced precision and recall for the MI class compared to the not-intoxicated (N) and low-intoxicated (L) classes (Table 6). The XGBoost-Mobile model exhibited a notably high false negative rate for instances labeled as "not-intoxicated," often misclassifying them as "moderate-intensive intoxicated."

However, it showed better accuracy in distinguishing "low-intoxicated" instances. In contrast, the XGBoost MobiFit model demonstrated a higher true positive rate compared to the other models, accurately identifying 76% of MI samples among the total samples belonging to that class. While the XGBoost-Mobile and Fitbit models achieved recall rates of 61% and 63% in predicting MI, they incorrectly predicted 56 and 53 out of 143 actual MI samples as other classes. In comparison, the best-performing MobiFit model achieved 108 true positives out of the 143 actual MI samples. The higher precision of the MobiFit model further supports its superior performance, though there remains room for improvement as it missed 35 samples, as shown in Table 6.

Table 5. Performance comparison of three XGBoost[a] models in detecting the subjective sense of moderate-intensive marijuana intoxication class.

ML[b] model	MI[c] precision	MI recall	MI F_1-score	MI AUC[d]
XGBoost-MobiFit	0.89	0.76	0.82	0.99
XGBoost-Mobile	0.64	0.61	0.62	0.96
XGBoost-Fitbit	0.65	0.63	0.64	0.98

[a]XGBoost: eXtreme Gradient Boosting.

[b]ML: machine learning

[c]MI: moderate-intensive intoxication.

[d]AUC: area under the curve.

Table 6. Confusion matrix for XGBoost-MobiFit, XGBoost-Mobile, and XGBoost-Fitbit model for 3 classes.

	Predicted		
	N[a]	L[b]	MI[c]
XGBoost[d] -MobiFit			
Actual			
N	6541	7	13
L	29	50	1
MI	35	0	108
XGBoost-Mobile			
Actual			
N	6452	59	50
L	28	52	0
MI	56	0	87
XGBoost-Fitbit			
Actual			
N	6499	14	48
L	41	39	0
MI	52	1	90

[a]N: not-intoxicated.

[b]L: low-intoxication.

[c]MI: moderate-intensive intoxication.

[d]XGBoost: eXtreme Gradient Boosting.

Key Features Contributing to Model Performance

Overview

To explore the algorithms' performance in predicting the MI class, we used SHAP summary visualizations [37,38] to identify patterns of acute marijuana intoxication. We determined the key features contributing significantly to the model's predictions based on mean absolute SHAP values across all instances, with a focus on the MI class.

Figures 5 and 6 present the SHAP visualizations. In Figure 5, the length of each bar on the left indicates the feature's contribution to the model, with longer bars signifying a stronger influence on the outcome. The SHAP summary plots on the right of Figure 5 illustrate how features influence the MI prediction class, with the strongest influence at the top. The

color shading indicates the direction of the feature's effect, with blue for low values, purple for median values, and red for high values. Plots extending to the left indicate a negative contribution to the prediction, while those extending to the right positively contribute to MI predictions.

Figure 5. Explanations generated by SHAP summary plot. Impact of features on best performing XGBoost-MobiFit model (left) and binary model output identifying moderate-intensive intoxication (MI; SHAP>0) from nonmoderate-intensive intoxication (N and L) classes (SHAP<0; right). HR: heart rate. SHAP: SHapley Additive exPlanations; WTSD: weighted stationary latitude and longitude standard deviation; XGBoost: eXtreme Gradient Boosting.

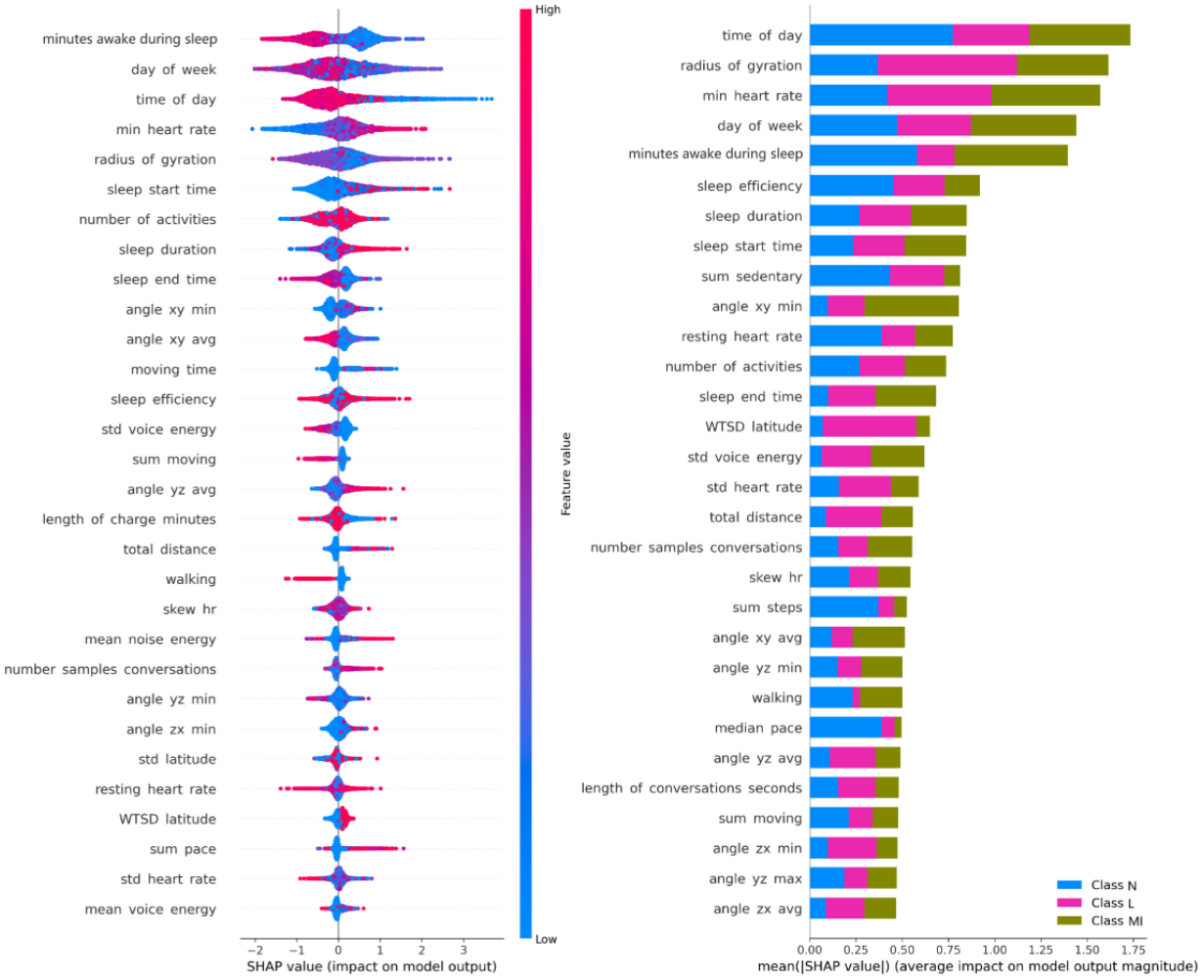
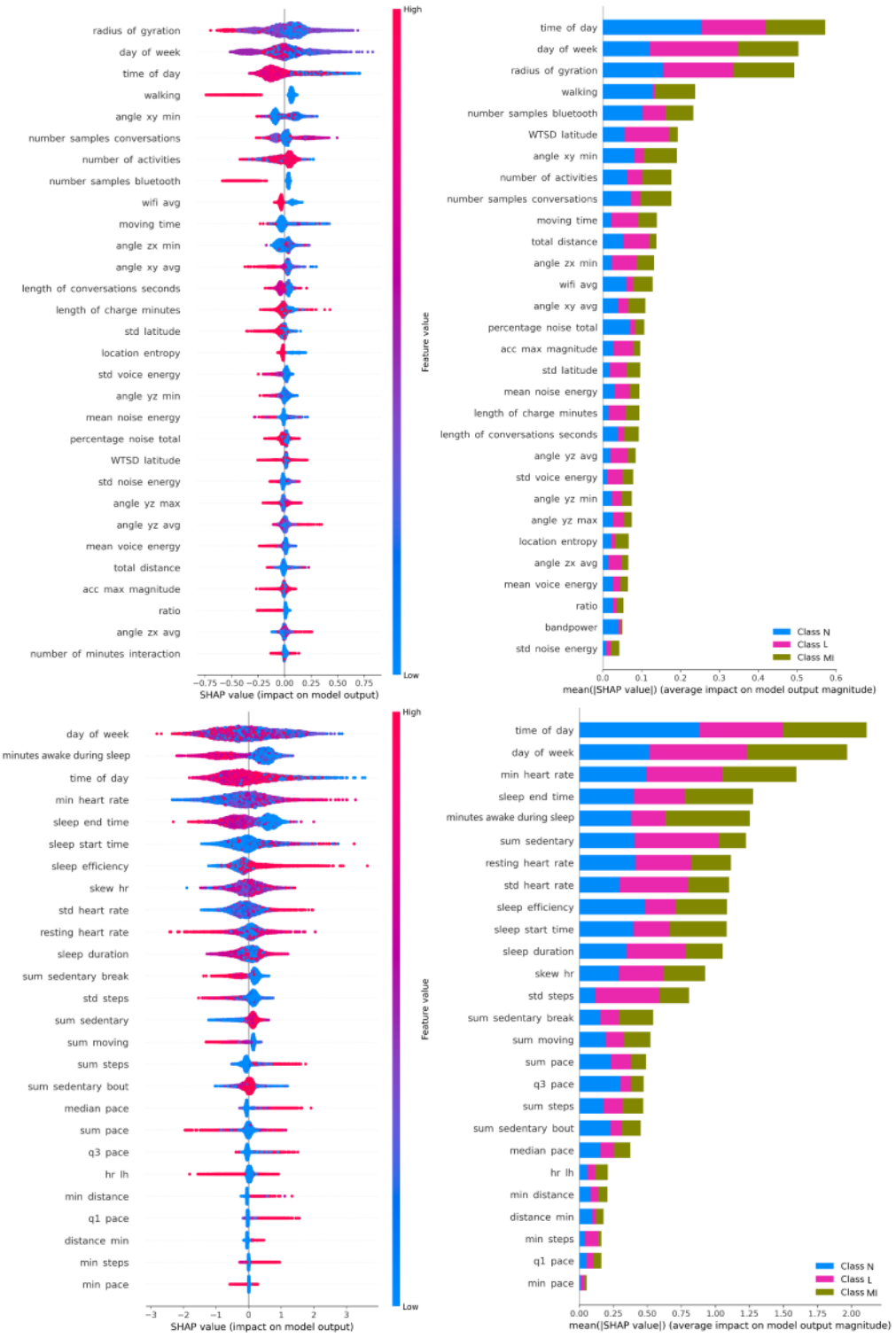


Figure 6. Explanations generated by SHAP summary plot. Impact of features on XGBoost-Mobile model (top left) and binary model output identifying MI (SHAP>0) from nonmoderate-intensive intoxication (N and L) classes (SHAP<0; top right), impact of features on XGBoost-Fitbit model (bottom left) and binary model output identifying MI (SHAP>0) from nonmoderate-intensive intoxication (N and L) classes (SHAP<0; bottom right). MI: moderate-intensive intoxication; SHAP: SHapley Additive exPlanations; WTSD: weighted stationary latitude and longitude standard deviation; XGBoost: eXtreme Gradient Boosting.



Impact of Average Key Features on Model Output Magnitude

The top five influential features in detecting the three classifications (Figure 5, left) and affecting the MI outputs (Figure 5, right) included time of day, radius of gyration, minimum HR, day of the week, and minutes awake during sleep. Among physical activities and physiological signals, a diverse range of features extracted from various sensors, including those beyond time-based attributes from both mobile and Fitbit combined sensors, was chosen as the top 30 crucial elements for distinguishing between not-intoxicated (N), low-intoxication

(L), and MI. The SHAP value, signifying the average impact magnitude on the model's output, played a pivotal role in this determination (Figure 5, left).

Impact of Unique Key Features on Mobile and Fitbit Model Outputs

Similar to the best-performing MobiFit model, the Mobile model (Figure 6) highlighted key features with overlapping impacts on the model's outcomes. The only exception was in specific movement and environmental context features, as shown in the top left and right graphs of Figure 6. However, the Fitbit model showed a more significant impact on HR features, with all four HR features ranking within the top 10 for all three classes (shown in the bottom-left graph in Figure 6), and for the MI classes compared to the non-MI classes (bottom-right graph in Figure 6).

Key Features Explaining MI

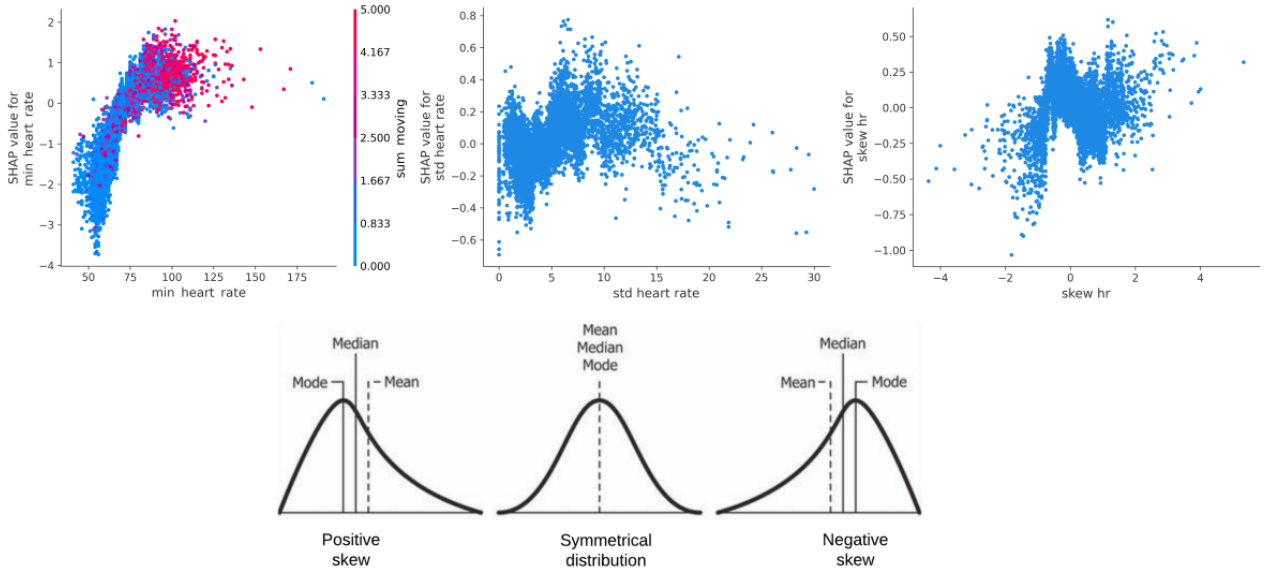
Overview

To specifically examine the influence of key features on the "risk" state of MI, we present comprehensive details for each key feature within the model.

A partial dependence plot (PDP) in Figure 7 illustrates the overall relationship between a feature and the outcome. The vertical axis represents SHAP values, signifying the effect of the chosen feature on predictions, while the horizontal axis represents actual feature values across instances. Each point represents an instance's feature value and its corresponding SHAP value. An upward PDP slope indicates a positive impact of the feature on MI prediction, while a downward slope indicates a negative impact. The surface on the PDP plot (eg, min HR and sum of moving minutes in Figure 7, top left) shows the combined impact of the two features on MI predictions, with greater values corresponding to increased prediction values.

In the following section, we introduce the key features contributing to MI, including elevated and fluctuating HR, reduced large-scale movement patterns, increased ambient noise and voice energy, and extended sleep patterns.

Figure 7. Interaction effects of total minutes spent moving on minimum HR values (top left), SD (top middle), and skewness (top right) of HR, and an explanation of skewness [39] (bottom). HR: heart rate; SHAP: SHapley Additive exPlanations.



Elevated and Fluctuating HRs

We investigated the impact of recent physical activity (measured as the sum of minutes spent moving based on Fitbit data) on HR in relation to self-reported marijuana intoxication using a PDP. The SHAP values for minimum HRs showed significant elevation, with an average increase from approximately 80 bpm to peaks of 90 bpm and reaching up to 100 bpm (ranging from 60 to 120 bpm, with a few data points exceeding 120 bpm). These elevated HRs corresponded to moderate-intensive self-reported marijuana intoxication (SHAP value>0) in young adults compared to other classes (not- and low-intoxicated).

The SHAP values clearly indicate *a* positive increase in minimum HR associated with a higher likelihood of self-reported MI, irrespective of the impact of the sum of minutes spent moving. The total movement time during self-reported MI influenced the rise in minimum HR, as shown in Figure 7 (top left), where the red values represent a maximum of 5 minutes of movement (our analysis uses 5-minute windows). While HR can fluctuate due to various factors, including physical activity, substance use (eg, alcohol), caffeine, meals, and mental state (eg, stress and anxiety), further research is needed to explore these additional influences.

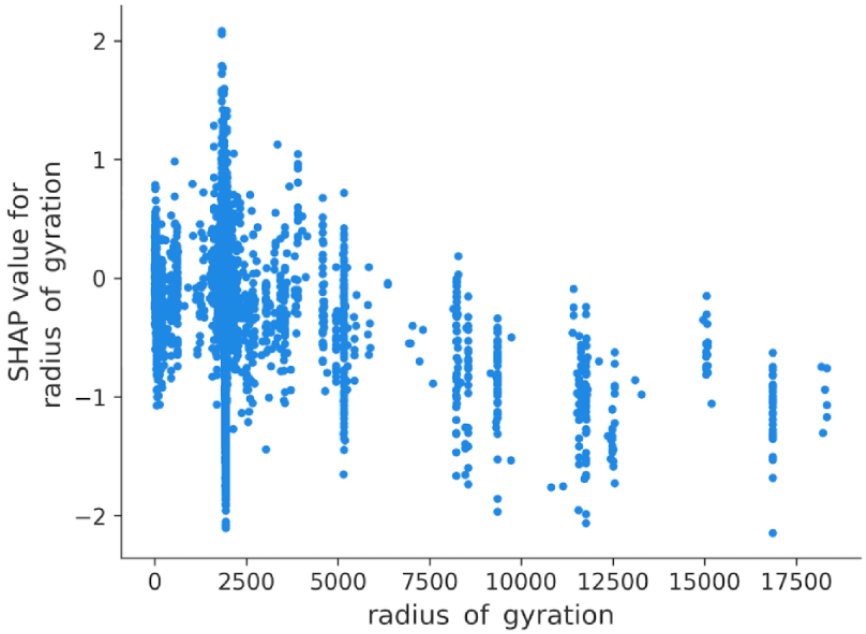
In brief, patterns for the SD of HRs exhibited fluctuations, but, in general, showed an increase when young adults reported MI (Figure 7, top middle). Negative skewness (indicating a "left-skewed" distribution) in HR was consistently associated with MI. This skewness suggests that there were more HR data

points on the right side of the mean (indicating that the median was greater than the mean), leading to a distribution stretched toward higher HR values (Figure 7, top right).

Decreased Large-Scale Movements

During MI, individuals showed a tendency for limited large-scale movement, often restricted to a radius of approximately 5 km. Notably, instances where the radius of gyration exceeded approximately 10 km were not associated with MI. This finding suggests that when young adults reported MI (rated 4-10), they were less inclined to engage in extensive travel (Figure 8). However, they still demonstrated movement within an average radius of 5 km.

Figure 8. Influence of radius of gyration (unit: meters). SHAP: SHapley Additive exPlanations.

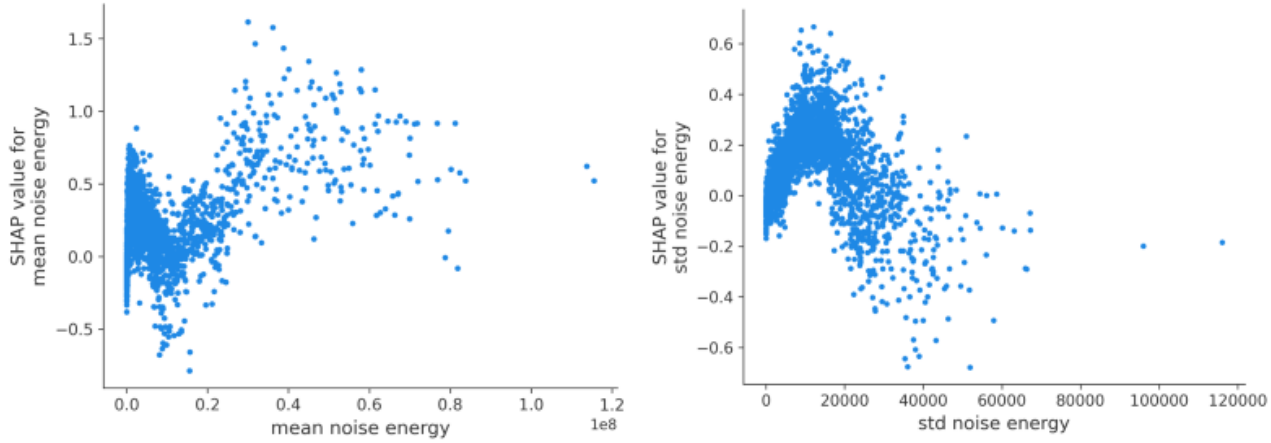


Elevated Surrounding Noise Energy

Interestingly, while the variance in environmental noise energy increased (with data points deviating further from the mean), the average noise energy decreased, though it exhibited an overall upward trend (Figure 9, left). Instances of MI were associated with increased noise variability (calculated based on the amplitude of audio samples), followed by a subsequent reduction (Figure 9, right).

Analyzing ambient sounds provides insights into the environmental context where individuals reporting MI might be located. This could include situations such as marijuana smoking, socializing with friends, or engaging with media like television or music. Although GPS-generated features were the primary indicators, MI may or may not be directly linked to specific locations such as shared social spaces (eg, lounges) or entertaining venues (eg, bars, pubs, or clubs). Nevertheless, it remains plausible that young adults reporting MI may choose to stay in noisy environments.

Figure 9. Influence of mean (left) and SD (right) noise energy (unit: Joule). SHAP: SHapley Additive exPlanations.



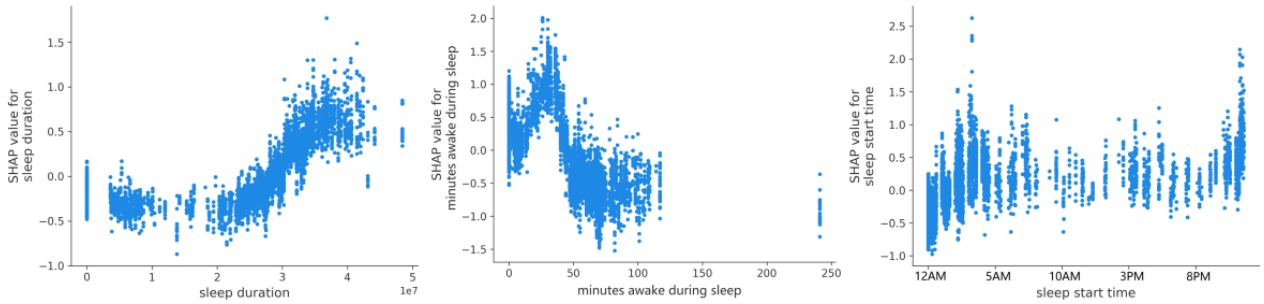
Prolonged Sleep Patterns

Distinct sleep patterns were linked to episodes of self-reported MI. Individuals who reported MI demonstrated extended sleep durations, spanning approximately 8 to 11 hours (Figure 10, left) the day before self-reported intoxication. In contrast, instances with low or no reported intoxication generally corresponded to healthy sleep durations, averaging around 6-7 hours, with some patterns as short as 2 hours.

There was also a positive correlation between the duration of minutes awake after falling asleep and self-reported MI, particularly when the period involved less than 50 minutes of wakefulness. However, an increase in extended minutes awake after falling asleep (if >50 minutes, extending beyond approximately an hour) did not show any significant association with a likelihood of MI (Figure 10, middle). Regarding sleep start times, the data indicated peaks at both 11 PM and early morning hours, with a rise in sleep start times continuing until around 4 AM (Figure 10, right).

In summary, elevated minimum HR values were clearly linked to a higher likelihood of self-reported MI. However, we observed that GPS-travel patterns (macromovements) did not appear to increase during self-reported marijuana intoxication. Interestingly, extended sleep hours and minutes awake during sleep [40] the day before self-reported marijuana intoxication were associated with MI.

Figure 10. Total sleep duration (left), minutes awake during sleep (middle), and sleep start time (right). SHAP: SHapley Additive exPlanations.



Additional Analyses for Real-World Feasibility

To enhance the practicality of our ML model in real-world settings, we conducted supplementary analyses to evaluate our top-performing model, the XGBoost-MobiFit model, under different scenarios. These scenarios involved: (1) excluding GPS-derived travel data due to potential privacy concerns or GPS deactivation; (2) excluding sleep data in cases where users did not provide sleep information; and (3) excluding both GPS-derived travel and sleep data. This approach aims to explore the feasibility of offering more flexible data collection options, potentially addressing privacy concerns and incomplete data issues.

In brief, excluding GPS-derived features (XGBoost-MobiFit-GPS excluded) resulted in a 15% decrease in the F_1-score compared to the best model, with a 10% reduction in sensitivity (recall). Excluding sleep data (XGBoost-MobiFit-Sleep excluded) led to a 24% decrease in the F_1-score compared to the best model. When both GPS and sleep features were excluded (XGBoost-MobiFit-GPS-Sleep excluded), the model experienced a 16% reduction in F_1-score and showed the lowest recall for identifying self-reported MI classes compared to the best-performing model. Please refer to Multimedia Appendix 5 for a detailed description of the additional analyses and results.

Discussion

Overview

The ability to detect subjective reports of acute marijuana intoxication in natural environments using mobile sensors has the potential to enable just-in-time interventions [41] to reduce marijuana-related harms. To the best of our knowledge, this is the first study that demonstrates the impact of integrating smartphone-based and wearable sensor features on the enhancement of the performance and interpretability of algorithms in detecting acute marijuana intoxication in naturalistic environments.

As hypothesized, we found that the XGB-MobiFit model, which combined smartphone sensor data with Fitbit features outperformed models that used only mobile or only Fitbit data. By integrating sensors from both smartphones and wearable devices, our best-performing algorithm balances specificity and sensitivity on unseen samples, enabling interpretable, transparent, and unobtrusive detection of acute subjective marijuana intoxication in natural environments. This opens up opportunities for real-time monitoring in everyday settings and the implementation of just-in-time adaptive interventions.

XAI visualizations supported our second hypothesis, highlighting HR, GPS, and physical movement data as key features that contributed to self-reported marijuana intoxication predictions. These findings were observed beyond the influences of simply applying time of day and day of the week features (ranked 1st and 4th, respectively), as validated in [11], particularly during instances of self-reported subjective marijuana intoxication in naturalistic environments.

Interpretable Behavioral and Physiological Signals of Marijuana Intoxication in Real-World Settings

To explain the results of the black-box ML models to detect marijuana intoxication in everyday settings, our study integrated sensors from smartphones and a wearable device, identified key sensor features, and used XAI to facilitate the interpretation of model results. The findings are consistent with prior research

conducted in controlled laboratory settings, which consistently found an acute increase in resting HR following marijuana use [12-14]. Our results suggest the potential for HR with behavioral factors to detect marijuana intoxication "outside of laboratory settings" using off-the-shelf devices in naturalistic environments. While many factors can affect HR in daily life, this study yielded significant HR features and insights from the elevated HR patterns during self-reported acute marijuana intoxication. Future research could explore associations between HR and other physiological and behavioral indicators of marijuana use, such as respiration, to better capture marijuana intoxication in natural environments [42].

The use of XAI visualization could help increase transparency and accountability when conducted as part of a substance use detection system [43, 44]. It is promising to use XAI as it enables researchers and clinicians to understand how algorithms arrive at decisions and identify key behavioral and physiological attributes, providing opportunities to improve detection accuracy and enhance trust in the algorithm over time.

Real-Time Detection and Intervention Potential

Compared to an average 30-minute marijuana episode, the 5-minute window used in the best-performing model is small enough to predict marijuana intoxication in near real-time. Detecting marijuana intoxication in near real-time promotes just-in-time intervention, which serves as a crucial first step toward reducing possible marijuana-related harm in a timely manner.

Our best detection model is unlikely to misclassify a "high" state as not high, which demonstrates the potential for using our detection algorithm with unseen data in real-world contexts. On the unseen test set, we obtained 85% precision (92% precision for 3 classes) in specifically identifying self-reported moderate-intensive marijuana intoxication. Passive sensing using smartphone-based sensors has been investigated in the context of alcohol intoxication [25,26,43], and here we extend this research to self-reported marijuana intoxication [11] beyond smartphone-based sensors, which could ultimately be useful for JIT interventions [41] to reduce marijuana-related harm. The value to society and individuals of reducing marijuana-related harm is clear. If individuals choose to use such a personal detection system, they will need to keep their phone charged and with them when using marijuana and wear a device (eg, Fitbit) and keep it charged as well.

For real-time modeling using the XGBoost algorithm, deploying the estimated model onto a computing device is an indispensable phase. We envisage two primary deployment scenarios: first, local assessments can be generated by deploying the model directly onto users' devices, such as smartphones. This approach ensures seamless functionality even without an internet connection but requires adequate storage and computational capacity. Second, cloud-based computation can be used. While this approach relies on a stable internet connection, it effectively offloads the computational burden from the user's device. Real-world applications introduce pragmatic considerations such as battery longevity, which could be affected by the model's continuous operation, and user privacy during data transmission and generation of model results.

Therefore, a comprehensive assessment of the model's feasibility in real-time operational settings is important. Our proposed generalized model, designed to operate across a diverse demographic spectrum rather than relying on individual-specific (idiographic) models, offers advantages in terms of scalability and practicality.

Privacy Considerations and User-Centric Configuration Choices

To highlight the benefits of combining sensor features from both smartphone and wearable devices while addressing potential privacy concerns, particularly related to location data, we aim to offer participants additional configuration choices rather than study withdrawal. For example, participants can deactivate GPS sensors if desired. This is demonstrated by our testing of the best-performing model, XGBoost-MobiFit, where we excluded location features. The analysis revealed a 15% (XGBoost-MobiFit-GPS excluded) decrease in F_1-score from the best model. As proposed by Bae et al [43], collecting GPS data and using rounded GPS data extraction (ie, less precise location data) could be a viable approach. This avoids using raw latitude and longitude, which may contain sensitive information on specific locations. Researchers and clinicians could consider providing alternative options instead of completely disabling GPS, as GPS data contributes to the model's accuracy.

Moreover, to assess the efficacy of our top-performing model, we conducted tests after excluding sleep-related features (Multimedia Appendix 5). The analysis revealed a 24% (XGBoost-MobiFit-Sleep excluded) decrease in the F_1-score compared to the best model's performance. While participants may benefit from the option to disable sensors when necessary, it is important to note that this could potentially decrease the model's ability to detect marijuana intoxication.

By building a system that prioritizes privacy and user autonomy, we can provide a valuable tool to reduce marijuana-related harm to individuals and society. Ultimately, each person will have to decide for themselves whether the benefits of a detection and intervention system outweigh the tradeoffs in minimizing possible marijuana-related harms to themselves and the broader community.

Limitations and Future Work

The first limitation of this study is relying on self-reporting as the ground truth, which may be subjective. This study extends prior ESM work, which codes self-reported marijuana use as yes or no [45], by asking participants to rate marijuana intoxication from 0 to 10, which may be subject to recall or other biases in reporting. The broad categorization might overlook nuanced differences within three categories: low-intoxication (1-3), moderate-intensive marijuana intoxication (4-10), and not-high (0), which could affect the accuracy of the classifiers. Future analyses examining the performance of mobile and wearable sensors against different thresholds for a subjective marijuana intoxication outcome could be valuable.

Another limitation was the size, diversity, and duration of the participants in the study. Since the participants were all young adults, the finding may not be generalizable to a broader age group. In addition, the level of compliance (63%) in completing the morning, afternoon, and evening surveys is relatively low. Thus, it is unclear whether all episodes of marijuana use were reported by participants, which could limit model performance. However, since there is no real-time accessible biological testing method at the time of publication, validating self-reported data with the current method still represents the best alternative. The current findings warrant future replication in a larger and more diverse group of participants over a longer period to address the limitations and validate the findings.

In addition, our model performed best when tested on the same participants it was trained on (with no overlap between training and testing data). While this has a valid use case, it assumes that we can always collect labeled training data for participants for whom we would like to apply the model. By applying more testing data, using more sophisticated sensor features, and better model tuning, future models could improve generalization over unseen testing participants. The HR data only holds significance when examined together with activity data. An acute increase in HR by itself is nonspecific and may not be associated with marijuana use or intoxication. False alarms triggered by the algorithm could erode trust in an automated system, whereas low sensitivity to actual marijuana use could result in marijuana-related harm. Therefore, it is important to investigate the interplay between human activities associated with marijuana intoxication and physiological signals in a larger population, and how these interactions can contribute to intervention delivery in real-world contexts.

Finally, it is crucial to acknowledge that the potential impact of polysubstance use on the interpretation of physiological signals associated with self-reported cannabis intoxication was not included. While ESM is used to collect information on the use of other substances, our analysis did not account for the effects of polysubstance use due to the limited scope of the study. The presence of polysubstance use could potentially confound the physiological signals attributed to marijuana. This may lead to inaccuracies in our algorithm, particularly in distinguishing between marijuana intoxication and the effects of other substances. Thus, while our study provides valuable insights into self-reported marijuana intoxication, it has limitations in addressing the full spectrum of real-world polysubstance use. Future research should include developing algorithms that can differentiate between the physiological signals associated with different substances, including polysubstance use.

Conclusions

Our study demonstrates that integrating features from smartphone-based sensors and wearable devices significantly improves the detection of self-reported marijuana intoxication in natural environments among young adults. The XGBoost-MobiFit model, which combines data from both smartphone sensors and wearable devices, achieved an F_1-score of 0.85 in detecting moderate to intensive self-reported marijuana intoxication, outperforming models that relied solely on smartphone sensors. The results suggest that incorporating wearable device data enhances the XGBoost model's performance by 13%, justifying the additional complexity of using wearable devices among young adults.

Key features contributing to the detection of self-reported "MI" included an acute increase in HR (measured by Fitbit), macromovement indicators (derived from GPS data), and prolonged sleep patterns the night before self-reported marijuana intoxication (measured by Fitbit).

Future research should focus on refining the algorithms that integrate smartphone and Fitbit sensor data in larger, more diverse samples. In addition, exploring how these algorithms, informed by XAI, can support the development of just-in-time interventions for clinicians is essential. Such interventions could offer context-adaptive, personalized strategies to minimize potential marijuana-related harms, such as intoxicated driving, therefore reducing the frequency and severity of acute marijuana-related incidents among young adults.

Acknowledgments

This study was supported by the National Institute on Drug Abuse (R21 DA043181/U01 DA056472), the Stevens Startup grant, and the Provost scholarship.

Authors' Contributions

SWB, TC, and AKD contributed to the design of the study and data collection. SWB, TZ, AKD, and RI processed the data, and SWB and TZ analyzed the data and developed the computational and explainable models. SWB drafted the initial manuscript, which was edited by TC, TZ, and AKD, and approved by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Extracted features overview.
[DOCX File , 40 KB - ai_v4i1e52270_app1.docx]

Multimedia Appendix 2
Feature selection.
[DOCX File , 197 KB - ai_v4i1e52270_app2.docx]

Multimedia Appendix 3
Rationale for machine learning model selection.
[DOCX File , 34 KB - ai_v4i1e52270_app3.docx]

Multimedia Appendix 4
Comparison of models with different classifiers.
[DOCX File , 41 KB - ai_v4i1e52270_app4.docx]

Multimedia Appendix 5
Privacy-preserving XGBoost-MobiFit models.
[DOCX File , 42 KB - ai_v4i1e52270_app5.docx]

References

1. Conroy DA, Kurth ME, Brower KJ, Strong DR, Stein MD. Impact of marijuana use on self-rated cognition in young adult men and women. Am J Addict 2015;24(2):160-165 [FREE Full text] [doi: 10.1111/ajad.12157] [Medline: 25864605]
2. Engineering National Academies of Sciences. The Health Effects of Cannabis and Cannabinoids: The Current State of Evidence and Recommendations for Research. Washington, DC: Academies Press; 2017.
3. Scott JC, Slomiak ST, Jones JD, Rosen AFG, Moore TM, Gur RC. Association of cannabis with cognitive functioning in adolescents and young adults: a systematic review and meta-analysis. JAMA Psychiatry 2018;75(6):585-595. [doi: 10.1001/jamapsychiatry.2018.0335] [Medline: 29710074]
4. Phillips KT, Phillips MM, Lalonde TL, Tormohlen KN. Marijuana use, craving, and academic motivation and performance among college students: an in-the-moment study. Addict Behav 2015;47:42-47 [FREE Full text] [doi: 10.1016/j.addbeh.2015.03.020] [Medline: 25864134]
5. Pertwee RG. Handbook of Cannabis. United Kingdom: Oxford University Press; 2015.
6. Ruojun LI, Emmanuel AGU, Ganesh B, Debra H, Ana A, Michael S. WeedGait: unobtrusive smartphone sensing of marijuana-induced gait impairment by fusing gait cycle segmentation and neural networks. 2019 Presented at: IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT); November 22, 2019; USA p. 94. [doi: 10.1109/hi-poct45284.2019.8962787]
7. Mishra RK, Sempionatto JR, Li Z, Brown C, Galdino NM, Shah R, et al. Simultaneous detection of salivary Δ-tetrahydrocannabinol and alcohol using a wearable electrochemical ring sensor. Talanta 2020;211:120757 [FREE Full text] [doi: 10.1016/j.talanta.2020.120757] [Medline: 32070607]
8. Pedersen ER, Hummer JF, Rinker DV, Traylor ZK, Neighbors C. Measuring protective behavioral strategies for marijuana use among young adults. J Stud Alcohol Drugs 2016;77(3):441-450. [doi: 10.15288/jsad.2016.77.441] [Medline: 27172576]
9. Huestis MA, Smith ML. Cannabinoid markers in biological fluids and tissues: revealing intake. Trends Mol Med 2018;24(2):156-172. [doi: 10.1016/j.molmed.2017.12.006] [Medline: 29398403]
10. Bédard M, Dubois S, Weaver B. The impact of cannabis on driving. Can J Public Health 2007;98(1):6-11. [doi: 10.1007/bf03405376]
11. Bae SW, Chung T, Islam R, Suffoletto B, Du J, Jang S, et al. Mobile phone sensor-based detection of subjective cannabis intoxication in young adults: a feasibility study in real-world settings. Drug Alcohol Depend 2021;228:108972-108716 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.108972] [Medline: 34530315]
12. Huber GL, Griffith DL, Langsjoen PM. The effects of marihuana on the respiratory and cardiovascular systems. Marijuana: An International Research Report. National Campaign Against Drug Abuse Monograph 7 (1988) 1988:123-134.
13. Maykut MO. Health consequences of acute and chronic marihuana use. Prog Neuro-Psychopharmacol Biol Psychiatry 1985;9(3):209-238. [doi: 10.1016/0278-5846(85)90085-5]
14. Zuurman L, Ippel AE, Moin E, van Gerven JMA. Biomarkers for the effects of cannabis and THC in healthy volunteers. Br J Clin Pharmacol 2009;67(1):5-21 [FREE Full text] [doi: 10.1111/j.1365-2125.2008.03329.x] [Medline: 19133057]
15. Carreiro S, Smelson D, Ranney M, Horvath KJ, Picard RW, Boudreaux ED, et al. Real-time mobile detection of drug use with wearable biosensors: a pilot study. J Med Toxicol 2015;11(1):73-79. [doi: 10.1007/s13181-014-0439-7] [Medline: 25330747]
16. Epstein DH, Tyburski M, Kowalczyk WJ, Burgess-Hull AJ, Phillips KA, Curtis BL, et al. Prediction of stress and drug craving ninety minutes in the future with passively collected GPS data. NPJ Digital Med 2020;3(1):26 [FREE Full text] [doi: 10.1038/s41746-020-0234-6] [Medline: 32195362]
17. Ferreira D, Kostakos V, Dey AK. AWARE: mobile context instrumentation framework. Front ICT 2015 Apr 20;2:1-9. [doi: 10.3389/fict.2015.00006]

18. Chung T, Bae SW, Mun E, Suffoletto B, Nishiyama Y, Jang S, et al. Mobile assessment of acute effects of marijuana on cognitive functioning in young adults: observational study. JMIR Mhealth Uhealth 2020;8(3):e16240 [FREE Full text] [doi: 10.2196/16240] [Medline: 32154789]
19. Mokrysz C, Freeman T, Korkki S, Griffiths K, Curran HV. Are adolescents more vulnerable to the harmful effects of cannabis than adults? A placebo-controlled study in human males. Transl Psychiatry 2016;6(11):e961 [FREE Full text] [doi: 10.1038/tp.2016.225] [Medline: 27898071]
20. Spindle TR, Cone EJ, Schlienz NJ, Mitchell JM, Bigelow GE, Flegel R, et al. Acute effects of smoked and vaporized cannabis in healthy adults who infrequently use cannabis: a crossover trial. JAMA Netw Open 2018;1(7):e184841 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.4841] [Medline: 30646391]
21. Mohr DC, Zhang MI, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol 2017;13:23-47 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
22. Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ 2016;4:e2537 [FREE Full text] [doi: 10.7717/peerj.2537] [Medline: 28344895]
23. Celebi ME, Celiker F, Kingravi HA. On Euclidean norm approximations. Pattern Recognit 2011;44(2):278-283. [doi: 10.1016/j.patcog.2010.08.028]
24. Denzil Ferreira. Com.Aware.Plugin.Studentlife.Audio_Final. Retrieved from GitHub. 2016. URL: https://github.com/denzilferreira/com.aware.plugin.studentlife.audio_final [accessed 2023-01-18]
25. Bae S, Chung T, Ferreira D, Dey AK, Suffoletto B. Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: implications for just-in-time adaptive interventions. Addict Behav 2018;83:42-47. [doi: 10.1016/j.addbeh.2017.11.039] [Medline: 29217132]
26. Bae S, Ferreira D, Suffoletto B, Puyana JC, Kurtz R, Chung T, et al. Detecting drinking episodes in young adults using smartphone-based sensors. Proc ACM Interact Mobile Wearable Ubiquitous Technol 2017;1(2):1-36. [doi: 10.1145/3090051] [Medline: 35146236]
27. Byrnes HF, Miller BA, Wiebe DJ, Morrison CN, Remer LG, Wiehe SE. Tracking adolescents with global positioning system-enabled cell phones to study contextual exposures and alcohol and marijuana use: a pilot study. J Adolesc Health 2015;57(2):245-247 [FREE Full text] [doi: 10.1016/j.jadohealth.2015.04.013] [Medline: 26206448]
28. Chaix B. Mobile sensing in environmental health and neighborhood research. Annu Rev Public Health 2018;39:367-384 [FREE Full text] [doi: 10.1146/annurev-publhealth-040617-013731] [Medline: 29608869]
29. Jensen MT, Suadicani P, Hein HO, Gyntelberg F. Elevated resting heart rate, physical fitness and all-cause mortality: a 16-year follow-up in the Copenhagen male study. Heart 2013;99(12):882-887 [FREE Full text] [doi: 10.1136/heartjnl-2012-303375] [Medline: 23595657]
30. American Heart Association. All about heart rate (pulse). Retrieved from. URL: https://www.heart.org/en/health-topics/high-blood-pressure/the-facts-about-high-blood-pressure/all-about-heart-rate-pulse [accessed 2024-11-05]
31. Tara K, Sarkar AK, Khan MAG, Mou JR. Detection of cardiac disorder using MATLAB based graphical user interface (GUI). 2017 Presented at: 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC); December 23, 2017; United States p. 440-443. [doi: 10.1109/r10-htc.2017.8288994]
32. Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesth Analg 2018;126(5):1763-1768. [doi: 10.1213/ANE.0000000000002864] [Medline: 29481436]
33. Yitzhaki S, Schechtman E. The Gini Methodology: A Primer on a Statistical Methodology. New York: Springer; 2013:11-31.
34. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-357. [doi: 10.1613/jair.953]
35. Takuya A, Shotaro S, Toshihiko Y, Takeru O, Masanori K. Optuna: a next-generation hyperparameter optimization framework. 2019 Presented at: KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; July 25, 2019; United States. [doi: 10.1145/3292500.3330701]
36. Powers D. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol 2011;2:37-63 [FREE Full text]
37. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56-67. [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
38. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30:4765-4774 [FREE Full text]
39. Skewness. 2023. URL: https://en.wikipedia.org/wiki/Skewness [accessed 2023-08-28]
40. Shrivastava D, Jung S, Saadat M, Sirohi R, Crewson K. How to interpret the results of a sleep study. J Community Hosp Intern Med Perspect 2014;4(5):24983. [doi: 10.3402/jchimp.v4.24983] [Medline: 25432643]
41. Joshua MS, Kristin EH. Is providing mobile interventions" just-in-time" helpful? An experimental proof of concept study of just-in-time intervention for stress management. 2016 Presented at: 2016 IEEE Wireless Health (WH); October 27, 2016; USA. [doi: 10.1109/wh.2016.7764561]
42. NIDA. What are marijuana's effects on other aspects of physical health?. URL: https://nida.nih.gov/research-topics/cannabis-marijuana [accessed 2023-08-10]

43. Bae SW, Suffoletto B, Zhang T, Chung T, Ozolcer M, Islam MR, et al. Leveraging mobile phone sensors, machine learning, and explainable artificial intelligence to predict imminent same-day binge-drinking events to support just-in-time adaptive interventions: algorithm development and validation study. JMIR Form Res 2023;7:e39862 [FREE Full text] [doi: 10.2196/39862] [Medline: 36809294]
44. Zhang T, Chung T, Dey A, Bae SW. Exploring Algorithmic Explainability: Generating Explainable AI Insights for Personalized Clinical Decision Support Focused on Cannabis Intoxication in Young Adults. 2024 Int Conf Act Behav Comput (2024) 2024 May;2024. [doi: 10.1109/abc61795.2024.10652070] [Medline: 39600343]
45. Randi MS, Robin J, Mermelstein, Donald H. Ecological momentary assessment of working memory under conditions of simultaneous marijuana and tobacco use. 2016. URL: https://doi.org/10.1111/add.13342

Abbreviations

AUC: area under the curve
bpm: beats per minute
ESM: experience sampling method
HR: heart rate
MI: moderate-intensive intoxication
ML: machine learning
PDP: partial dependence plot
SHAP: SHapley Additive exPlanations
THC: delta-9 tetrahydrocannabinol
XAI: explainable artificial intelligence
XGBoost: eXtreme Gradient Boosting

Edited by K El Emam, B Malin; submitted 29.08.23; peer-reviewed by E Karoulla, Q Liu, I Liu; comments to author 09.11.23; revised version received 31.01.24; accepted 02.09.24; published 02.01.25.

Please cite as:
Bae SW, Chung T, Zhang T, Dey AK, Islam R
Enhancing Interpretable, Transparent, and Unobtrusive Detection of Acute Marijuana Intoxication in Natural Environments: Harnessing Smart Devices and Explainable AI to Empower Just-In-Time Adaptive Interventions: Longitudinal Observational Study
JMIR AI 2025;4:e52270
URL: https://ai.jmir.org/2025/1/e52270
doi:10.2196/52270
PMID:

Original Paper

Urgency Prediction for Medical Laboratory Tests Through Optimal Sparse Decision Tree: Case Study With Echocardiograms

Yiqun Jiang[1], PhD; Qing Li[2], PhD; Yu-Li Huang[1], PhD; Wenli Zhang[3], PhD

[1]Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN, United States

[2]Department of Industrial & Manufacturing Systems Engineering, Iowa State University, Ames, IA, United States

[3]Department of Information Systems and Business Analytics, Iowa State University, Ames, IA, United States

Corresponding Author:
Wenli Zhang, PhD
Department of Information Systems and Business Analytics
Iowa State University
2167 Union Drive
Ames, IA, 50011-2027
United States
Phone: 1 5152942469
Email: wlzhang@iastate.edu

Abstract

Background: In the contemporary realm of health care, laboratory tests stand as cornerstone components, driving the advancement of precision medicine. These tests offer intricate insights into a variety of medical conditions, thereby facilitating diagnosis, prognosis, and treatments. However, the accessibility of certain tests is hindered by factors such as high costs, a shortage of specialized personnel, or geographic disparities, posing obstacles to achieving equitable health care. For example, an echocardiogram is a type of laboratory test that is extremely important and not easily accessible. The increasing demand for echocardiograms underscores the imperative for more efficient scheduling protocols. Despite this pressing need, limited research has been conducted in this area.

Objective: The study aims to develop an interpretable machine learning model for determining the urgency of patients requiring echocardiograms, thereby aiding in the prioritization of scheduling procedures. Furthermore, this study aims to glean insights into the pivotal attributes influencing the prioritization of echocardiogram appointments, leveraging the high interpretability of the machine learning model.

Methods: Empirical and predictive analyses have been conducted to assess the urgency of patients based on a large real-world echocardiogram appointment dataset (ie, 34,293 appointments) sourced from electronic health records encompassing administrative information, referral diagnosis, and underlying patient conditions. We used a state-of-the-art interpretable machine learning algorithm, the optimal sparse decision tree (OSDT), renowned for its high accuracy and interpretability, to investigate the attributes pertinent to echocardiogram appointments.

Results: The method demonstrated satisfactory performance (F_1-score=36.18% with an improvement of 1.7% and F_2-score=28.18% with an improvement of 0.79% by the best-performing baseline model) in comparison to the best-performing baseline model. Moreover, due to its high interpretability, the results provide valuable medical insights regarding the identification of urgent patients for tests through the extraction of decision rules from the OSDT model.

Conclusions: The method demonstrated state-of-the-art predictive performance, affirming its effectiveness. Furthermore, we validate the decision rules derived from the OSDT model by comparing them with established medical knowledge. These interpretable results (eg, attribute importance and decision rules from the OSDT model) underscore the potential of our approach in prioritizing patient urgency for echocardiogram appointments and can be extended to prioritize other laboratory test appointments using electronic health record data.

(JMIR AI 2025;4:e64188) doi:10.2196/64188

KEYWORDS

interpretable machine learning; urgency prediction; appointment scheduling; echocardiogram; health care management

Introduction

Background

In the present medical landscape, the intricate interplay between innovative techniques has expanded the horizons of medical knowledge and opened avenues for unprecedented precision in patient care. The increasingly sophisticated laboratory tests play a crucial role in this transformative process. Born out of meticulous research and honed by the rigors of scientific scrutiny, these tests provide clinicians with a multifaceted toolkit to decipher the intricacies of illnesses, capturing the nuances of each condition, guiding medical professionals toward evidence-based interventions, and empowering medical professionals to tailor treatments with personalized precision.

However, a pivotal factor to take into consideration is the limited availability of certain state-of-the-art laboratory tests, as they often involve intricate equipment and elaborate protocols. This is evident from their expensive nature, the scarcity of skilled medical professionals capable of operating these laboratories, and the limited accessibility across different regions or during specific time frames [1]. As a result, the transformative potential of these laboratory tests is mitigated by the practical challenges they pose in terms of affordability [2]. The potential significant advantages of laboratory tests, coupled with their limited availability, render them a scarce resource, resulting in many patients having to endure wait times for access to laboratory tests. Consequently, predicting and prioritizing which patients require testing has emerged as an important research problem.

The rise of health IT and the subsequent influx of electronic health record (EHR) data, combined with the power of machine learning, offers new opportunities to revolutionize the prioritization of medical laboratory tests [3]. By delving into vast amounts of historical patient information, machine learning algorithms can discern intricate patterns and correlations that might otherwise elude human observation. The predictive outcomes generated by machine learning algorithms can contribute to refining testing protocols, enabling medical practitioners to make data-driven decisions regarding the prioritization and scheduling of laboratory tests based on patient information. In this study, we aim to elucidate methods for evaluating patients' urgency for tests, seeking to refine the allocation of scarce laboratory tests by harnessing the power of machine learning and analyzing historical EHRs. Specifically, we aim to contribute by applying an optimal sparse decision tree (OSDT) to a new domain—predicting the urgency of medical laboratory tests, using echocardiograms as a case study. Based on our literature review, OSDT stands out as one of the most suitable methods for achieving both optimal performance and interpretability in predicting the urgency of patients requiring echocardiograms. Our ultimate objective is to ensure prompt access for patients with the most critical needs.

Related Work

Echocardiogram and Patient Prioritization Techniques

An echocardiogram is one the most cost-effective means for screening cardiac anatomy, uses ultrasound to evaluate the cardiac structures, and provides critical information for medical providers [4]. It functions as a crucial precursor to a detailed diagnosis, capable of screening cardiac anatomy and providing essential information for assessing cardiovascular conditions such as murmurs, stenosis, and regurgitation. Additionally, it plays a crucial role in diagnosing valvular morphology and uncovering the root causes of valve diseases [5]. A comprehensive echocardiographic assessment can provide both diagnostic and prognostic information, thus facilitating risk stratification and establishing baseline data for future evaluations [5].

The echocardiogram, although immensely valuable, is not always easily attainable due to the increasing demand for the test. For example, there has been an observed increase in the prevalence of rheumatic heart disease, which stands as the most predominant form of valvular heart disease and impacts approximately 41 million individuals in developing countries [6]. In recent years, there has been a notable escalation in the demand for pediatric cardiology services, leading to documented workloads that have exhibited a substantial upsurge of up to 51% over the past decades [7]. Furthermore, there has been an increase in the prevalence of children with asymptomatic murmurs who necessitate evaluation through echocardiogram [8]. The increasing demands pose challenges to echocardiogram laboratories in resource management, requiring medical institutions to establish more effective scheduling protocols to prioritize patients in critical need of echocardiogram lab appointments.

Patient prioritization techniques can be broadly classified into scoring systems and machine learning classification–based systems [9]. Scoring systems, particularly those using regression techniques, have gained prominence for their ability to allocate medical resources. These systems heavily rely on the expertise of medical professionals to assign priority scores to patients. Examples include the Salisbury priority scoring system, allowing surgeons to assign relative priorities, and the Italian waiting time prioritization system, which reallocates outpatient referrals based on clinical priorities prescribed by general practitioners [9]. These methods, however, exhibit various limitations. First, there may be inherent bias (eg, subjective judgments obtained through experience by medical professionals) as these approaches often necessitate input from medical specialists' judgments. A machine learning and data-driven method can serve as a complement to these types of systems. Second, these methods might be tailored for a particular patient prioritization task (eg, surgery or referral), and demand a high level of specialized medical knowledge for their design, making them difficult to generalize to other tasks [10]. Third, certain methods lack transparent decision rules for assessing the significance of input attributes, thereby posing challenges for their practical applications [11]. Machine learning classification-based methods typically rely on a large amount of patients' information (eg, EHRs) to autonomously discern patterns and generate predictions. This process aids in patient prioritization and avoids limitations associated with scoring systems [12]. The existing methods, however, fail to transform the prediction process and outcomes into clear and executable rules, limiting the practical application of these approaches [9]. Moreover, existing studies predominantly center around 5 clinical areas, including cataract

surgery, general surgical procedures, hip and knee replacements, magnetic resonance imaging scanning, and children's mental health using specific predictive attributes and expert systems [13]. There is a crucial need for new methods that apply more broadly to general laboratory test prioritization.

To summarize, our literature review underscores the need for new methods of prioritizing patients, which leverage machine learning and data-driven techniques to complement existing methods, ensure transparency, and have the potential to be generalized to various patient prioritization tasks. Consequently, using extensive patient historical EHRs combined with an interpretable machine learning approach emerges as a potential solution to address these gaps.

Leveraging Machine Learning for Optimizing the Use of Scarce Laboratories Tests

When a large number of patient EHRs, which contain numerous hidden patterns, are available, integrating machine learning into health care practices emerges as a potential solution to address pressing issues such as the continual demand for medical services outpacing available resources. Specifically, machine learning, with its capacity to analyze vast data and discern intricate patterns, empowers health care professionals to make data-driven decisions regarding the allocation of laboratory tests. By developing predictive models using historical EHRs, machine learning models can identify individuals who are more likely to benefit from specific tests, ensuring that scarce resources are allocated where they can yield the greatest impact. Furthermore, such methods ensure critical cases receive prompt attention, leading to expedited diagnoses and interventions [14]. Moreover, the prediction results can potentially streamline the testing process by reducing unnecessary tests [15].

The integration of machine learning techniques to optimize the allocation of limited medical tests and laboratory resources has attracted considerable research attention. Research by Elitzur et al [16] delves into the use of prediction models to allocate medical tests efficiently. The study uses historical patient data to develop models that identify the most suitable candidates for specific tests, thereby enhancing resource allocation and streamlining the testing process. In a similar vein, Marescotti et al [17] investigate the orchestration of laboratory workflows through machine learning-driven prioritization. By considering factors such as clinical urgency and resource availability, their work demonstrates how machine learning algorithms can ensure timely and effective laboratory test processing, contributing to both improved patient care and optimized resource use. Similarly, Zhang et al [18] estimate the probability of requiring mechanical ventilation for in-hospital patients and contribute to the literature by identifying which patients require medical devices (ie, critical medical resources) more urgently.

However, while the potential benefits of machine learning in optimizing resource allocation are evident, challenges remain. A recent study underscores the need for further research and development in the area of machine learning models' interpretability and fairness, ensuring that data-driven decisions in health care maintain transparency [19]. The research gap drives us to use an interpretable and efficient machine learning method for laboratory tests and patient optimization.

Interpretable Machine Learning

Medical research is often at the forefront of technological innovation, with machine learning algorithms being harnessed to analyze vast datasets, predict disease outcomes, and assist in clinical decision-making. However, as these algorithms become increasingly sophisticated, they tend to function as "black boxes," where the reasoning behind their predictions remains obscured. This opacity not only raises concerns about trustworthiness but also impedes the adoption and acceptance of these tools by medical professionals [19].

In medical research, the concept of interpretability holds profound significance. The intricate interplay between cutting-edge technology and human well-being underscores the critical need to not only generate accurate predictions but also to understand the underlying rationale behind those predictions. The complexity of medical data, coupled with the potential life-altering consequences of decisions made based on data and machine learning models, demands a heightened level of transparency and comprehensibility requirements [20].

The interpretability of machine learning models empowers health care providers to understand the factors that led to a specific decision, enabling them to fine-tune treatment strategies according to their medical judgment and the patient's unique circumstances. Consequently, there has been a surge in post hoc techniques for elucidating black box machine learning models in a manner interpretable by humans. The most prominent techniques among these include local, model-agnostic methods that aim to explain individual predictions of a given black box classifier, such as local interpretable model-agnostic Explanation and Shapley additive explanation [21]. Due to their high generalizability, post hoc methods have been used to explain a wide array of machine learning models across various domains. However, previous research has indicated that there are common limitations associated with these post hoc techniques, including local interpretability, sensitivity to perturbations, and difficulties in choosing interpretable surrogate models [21].

In health care, arguably, a more appropriate research direction for using interpretable machine learning is tree-based models because much of the data related to patient prioritization is structured data (eg, tabular EHRs). Tree-based machine learning models can perform comparably to complex models (eg, deep learning models), especially after thorough preprocessing of tabular data [22]. In contrast to post hoc explainable machine learning techniques, tree-based models are logical models that consist of statements involving logical operations, providing clear and interpretable decision rules [22]. This interpretability is highly valuable in health care, as it allows medical professionals to not only make accurate predictions but also understand the underlying factors driving those predictions, enhancing transparency and trust in the decision-making process.

Since our research aims to use historical EHR data for patient prioritization, it is crucial to acknowledge another notable characteristic of patient prioritization-related information: the prevalence of numerous categorical variables (eg, patient demographic information such as gender and age groups). Furthermore, the outcomes of patient prioritization are also

expressed as categorical variables. For example, preventive interventions often involve categorical decisions, such as determining which individuals should undergo selective or indicated interventions or identifying those most likely to benefit from specific treatments [23]. In such scenarios, an efficient tree-based approach tailored to categorical variables is highly valuable. In this study, we focus on a cutting-edge decision tree algorithm–OSDT [24].

A decision tree features a hierarchical structure that is composed of a root node, branches, internal nodes, and leaf nodes in a tree format. Each path from the root node to the leaf node illustrates a rule to partition the data and leads to the final classification. The tree-based method presents a clear pattern for the decision-making process; thus, it is considered a transparent and highly interpretable model [25]. The results of the tree-based models are extremely useful for medical decision-making [26], and the performance of decision tree classifiers is verified by researchers on medical data [27]. Nevertheless, concerns have been raised regarding the suboptimality of decision tree algorithms [24,28]. To address this issue, OSDT has been introduced, aiming to ensure optimal solutions for binary variables in a computationally efficient manner [24].

The OSDT algorithm addresses various limitations observed in prior tree-based methods. Unlike previous approaches that often focused on finding the optimal tree within a fixed number of nodes or limited topology, OSDT tackles these shortcomings by identifying optimal trees through the use of a regularized loss function. This loss function strikes a balance between accuracy and the number of leaves, thereby enhancing the efficiency of the decision tree model. Furthermore, OSDT improves computational efficiency and interpretability by incorporating a series of analytical bounds that effectively reduce the search space while still identifying the optimal tree. By implementing these bounds, the algorithm streamlines the search process, leading to expedited identification of the optimal decision tree structure. Moreover, the OSDT algorithm has undergone mathematical validation, demonstrating its efficacy in constructing optimal trees for structured tabular datasets with attributes having binary values. It establishes its effectiveness in addressing binary classification problems. The algorithm is designed to uphold commendable levels of accuracy and is anticipated to meet the demands of medical prediction tasks with stringent interpretability requirements.

Methods

Study Design

In this study, we conducted empirical and predictive analyses using echocardiogram data extracted from EHRs at a large multispecialty hospital and medical facility. The dataset included administrative details, referral diagnoses, and patient conditions. To explore attributes relevant to echocardiogram prioritization, we used the OSDT algorithm due to its high accuracy and interpretability. We aim to enhance the scheduling of echocardiogram laboratory appointments by enabling the prioritization of patients with urgent needs based on our model's predictions. To be noted, our proposed method is not intended to replace human expertise but to complement it, offering valuable insights that guide practitioners toward informed and patient-centric choices.

Ethical Considerations

The Mayo Clinic Institutional Review Board, based on the authors' submission notes and in accordance with the Code of Federal Regulations, 45 CFR 46.102, deemed that this research did not require IRB review.

Data Collection and Selection

The dataset comprises real-world data from one of the top medical centers in the United States. The data were collected over a 1-year period in 2019, including 34,293 echocardiogram appointments. It consisted of 64 dummy-coded categorical attributes, encompassing various aspects such as patient demographics, medical history, clinical settings (eg, inpatient or outpatient status), past procedures, future scheduled procedures, and diagnose indicators for echocardiogram-justifying signs (eg, heart murmurs, shortness of breath, or chest pain) extracted from the clinical notes and referrals in the EHRs (Table 1).

The dataset exhibited a notable class imbalance issue, particularly evident in the examination of the "MadeBeforeEcho" attribute. This attribute delineates whether the downstream appointment following the echocardiogram occurs before the scheduling date of the echocardiogram appointment (not the actual appointment date). Within the "Y" category, the distribution revealed 84% nonurgent cases and 16% urgent cases. Conversely, in the "N" category, the distribution portrayed 58% nonurgent cases and 42% urgent cases. This observation underscored a substantial prevalence of nonurgent cases within the "MadeBeforeEcho" attribute. Furthermore, a similar pattern of imbalance is discerned when analyzing attributes such as "ReferredType" and "SurgeryYN." These attributes also exhibit a significant majority of cases concentrated within 1 category, indicating the need for careful consideration of class distribution in subsequent predictions.

The response variable is determined by calculating the number of days between the date the echocardiogram appointment was generated in the system and the actual appointment date. According to medical policy, appointments are classified as urgent (ie, the response variable) if the number of days is 2 or less, and nonurgent otherwise.

It is important to note that the features categorized under the "Future Scheduled Process" were derived based on the date the echocardiogram appointment is generated in the system, rather than the actual appointment date (Figure 1). This approach ensures that the model uses only the information available up to the point of echocardiogram appointment generation, without incorporating any data beyond this cutoff.

Of note, our dataset is a tabular dataset with attributes and response variables having binary values. Therefore, OSDT is highly suitable for serving this dataset, assisting us in making predictions for patient prioritization.

Table 1. Dataset and attribute statistics[a].

Category and variable	Description	Summary statistics, n (%)	
		Nonurgent	Urgent
Demographics			
Age (years)			
0-18	—[b]	1929 (7.18)	478 (6.41)
19-55	—	6766 (25.19)	1930 (25.90)
56-65	—	4954 (18.45)	1342 (18.01)
66-75	—	6784 (25.26)	1896 (25.44)
Older than 75	—	6398 (23.82)	1775 (23.82)
Sex			
Female	—	11,829 (44.09)	3529 (47.55)
Male	—	15,002 (55.91)	3892 (52.45)
Patient geolocation			
In_State	—	9973 (37.14)	2376 (31.96)
Out_of_State	—	14,332 (53.37)	4301 (57.85)
Town	—	2550 (9.50)	758 (10.20)
Clinical settings			
ReferralType			
External	—	1156 (4.30)	606 (8.15)
Internal	—	25,699 (95.70)	6829 (91.85)
ReferredBy	The specialty that patient referred by		
Cardiovascular medicine	—	8188 (30.49)	1162 (15.63)
Family medicine	—	436 (1.62)	142 (1.91)
Hospital medicine	—	145 (0.54)	4 (0.05)
Internal medicine	—	978 (3.64)	591 (7.95)
Obstetrics and gynecology	—	1096 (4.08)	359 (4.83)
Pediatric and adolescent medicine	—	2302 (8.57)	401 (5.39)
Other	—	13,710 (51.05)	4776 (64.24)
ReferredFrom	Referral origin		
Arizona campus	—	2 (0.01)	0 (0.00)
Florida campus	—	1 (0.00)	0 (0.00)
Mayo Clinic health system	—	154 (0.57)	38 (0.51)
Rochester campus	—	17,495 (65.15)	4463 (60.03)
Other	—	9203 (34.27)	2934 (39.46)
ReferredType	Referred type		
Outpatient	—	18,706 (69.66)	4585 (61.52)
Other	—	8149 (30.34)	2868 (38.48)
Future scheduled process			
Diff_surgery_after	The number of days between the date the echocardiogram appointment was generated in the system and the surgery date		
0-1	—	1449 (5.40)	461 (6.20)
2-5	—	1607 (5.98)	492 (6.62)

Category and variable	Description	Summary statistics, n (%)	
		Nonurgent	Urgent
6-15	—	1143 (4.26)	606 (8.15)
16 and greater	—	4715 (17.56)	1494 (20.09)
None	—	17,941 (66.81)	4382 (58.94)
MadeBeforeEcho	Whether the next downstream appointment after echocardiogram is made before the date the echocardiogram appointment was generated in the system or not		
Yes	—	23,845 (88.79)	4660 (62.53)
No	—	3010 (11.21)	2793 (37.47)
NextDepartment	The department in which the appointment happened after the date the echocardiogram appointment was generated in the system		
Cardiovascular medicine	—	12,012 (44.73)	1749 (23.47)
Non-cardiovascular medicine	Departments other than cardiovascular medicine	14,843 (55.27)	5704 (76.53)
NextLength	The number of days from the date the echocardiogram appointment was generated in the system to its following appointment		
0-1	—	4531 (16.87)	1608 (21.63)
1-5	—	3301 (12.29)	2018 (27.14)
Greater than 5	—	1,014 (3.78)	618 (8.31)
None	—	18,009 (67.06)	3191 (42.92)
Procedure	Type of echocardiogram visit		
TEE[c]	—	848 (3.16)	362 (4.87)
TTE[d]	—	23,293 (86.74)	6803 (91.50)
Other	—	2714 (10.11)	270 (3.63)
Past procedures			
SurgeryYN	Whether the patient had a cardiovascular surgery within 6 months prior to the date the echocardiogram appointment was generated in the system		
Yes	—	1708 (6.36)	264 (3.54)
No	—	25,147 (93.64)	7189 (96.46)
SurgeryYN_After	Whether the patient had a surgery within 3 months after the date the echocardiogram appointment was generated in the system		
Yes	—	8914 (33.19)	3053 (40.96)
No	—	17,941 (66.81)	4400 (59.04)
Medical history			
Alcohol	Alcohol abuse	115 (0.43)	50 (0.67)
Anemia	Anemia	962 (3.58)	605 (8.12)
BloodLoss	Blood loss	87 (0.32)	33 (0.44)
CHF[e]	—	1884 (7.02)	484 (6.49)
Coagulopathy	Coagulation deficiency	446 (1.66)	274 (3.68)
Depression	Major depressive disorder	439 (1.63)	192 (2.58)
DM[f]	—	610 (2.27)	230 (3.09)

Category and variable	Description	Summary statistics, n (%)	
		Nonurgent	Urgent
DMcx[g]	—	317 (1.18)	129 (1.73)
Drugs	Drug abuse	86 (0.32)	19 (0.25)
FluidsLytes	Fluid and electrolyte disorders	1013 (3.77)	617 (8.28)
HIV	—	0 (0.00)	1 (0.01)
Hypertension	—	2201 (8.20)	786 (10.55)
Hypothyroid	Hypothyroidism	777 (2.89)	277 (3.72)
Liver	—	429 (1.60)	197 (2.64)
Lymphoma	Lymph system cancer	464 (1.73)	347 (4.66)
Metastatic cancer	—	251 (0.93)	222 (2.98)
NeuroOther	Neurological disorders	581 (2.16)	291 (3.90)
Obesity	—	980 (3.65)	339 (4.55)
Paralysis	—	58 (0.22)	15 (0.20)
PHTN[h]	Pulmonary circulation disorders	298 (1.11)	153 (2.05)
Psychoses	Mental disorder characterized by a disconnection from reality	126 (0.47)	53 (0.71)
PUD[i]	Chronic peptic ulcer	41 (0.15)	20 (0.27)
Pulmonary	Chronic pulmonary disease	650 (2.42)	273 (3.66)
PVD[j]	—	965 (3.59)	234 (3.14)
Renal	Renal failure	950 (3.54)	331 (4.44)
Rheumatic	Rheumatoid arthritis or collagen vascular	254 (0.95)	150 (2.01)
Tumor	Solid tumor	722 (2.69)	380 (5.10)
Valvular	Valvular disease	3367 (12.54)	573 (7.69)
WeightLoss	Weight loss	248 (0.92)	237 (3.18)
Diagnoses			
A	MSSA[k] bacteremia, sepsis	18 (0.07)	25 (0.34)
B	MRSA[l], staph bacteremia, slaph, fungemia, pseudomonas, candidemia, MRSA bacteremia	47 (0.18)	40 (0.54)
C	Leukemia, AML[m], CML[n], lymphoma, AMV[o], myeloma	1428 (5.32)	554 (7.43)
D	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	561 (2.09)	193 (2.59)
E	Endocrine, nutritional and metabolic diseases	1714 (6.38)	408 (5.74)
F	Behavioral and neurodevelopmental disorders	49 (0.18)	46 (0.62)
G	Muscular dystrophy	590 (2.20)	273 (3.66)
H	Diseases of the eye and adnexa or disease of the ear and mastoid process	60 (0.22)	28 (0.38)
I	Heart failure, coronary artery, cardiac arrest, STEMI[p], stroke, cardia, hypertension, endocarditis, NSTEMI[q], PEA[r] arrest, AFib[s], pulmonary embolism, pulmonary hypertension, and vegetation	11,302 (42.09)	4096 (54.96)

Category and variable	Description	Summary statistics, n (%)	
		Nonurgent	Urgent
J	Resp failure, respiratory, and pulmonary	477 (1.78)	392 (5.26)
K	Liver and cirrhosis	357 (1.33)	130 (1.74)
L	Diseases of the skin and subcutaneous tissue	36 (0.13)	33 (0.44)
M	Diseases of the musculoskeletal system and connective tissue	503 (1.87)	280 (3.76)
N	Diseases of the genitourinary system	397 (1.48)	119 (1.60)
O	Pre-eclampsia, preeclampsia	235 (0.88)	57 (0.76)
P	Certain conditions originating in the perinatal period	12 (0.04)	4 (0.05)
Q	Ehlers, coarc, PDA[t], and congenital	2811 (10.47)	309 (4.15)
R	Murmur, hypoxemia, shortness, SOB[u], breath, shock, dyspnea, chest pain, troponin, syncope, electrocardiogram, extremity, mass, and swelling, edema	4111 (15.31)	2811 (37.72)
S	Injury, poisoning and certain other consequences of external causes	100 (0.37)	21 (0.28)
Z	Chemo, preoperative, pre-op, prenatal, pregnancy, prior to, BMI, surgery, and transplant	5966 (22.22)	1129 (15.15)

[a]All the features used in this study are complete for each patient, with no missing values. The diagnoses are derived from patients' ICD-9 codes, and the medical history is extracted from electronic health record notes using the medical center's built-in natural language processing tools.

[b]Not applicable.

[c]TEE: transesophageal echocardiogram.

[d]TTE: transthoracic echocardiogram.

[e]CHF: congestive heart failure.

[f]DM: diabetes without chronic complications.

[g]DMcx: diabetes with chronic complications.

[h]PHTN: pulmonary hypertension.

[i]PUD: peptic ulcer disease.

[j]PVD: peripheral vascular disease.

[k]MSSA: methicillin-sensitive *Staphylococcus aureus*.

[l]MRSA: methicillin-resistant *Staphylococcus aureus*.

[m]AML: acute myeloid leukemia.

[n]CML: chronic myeloid leukemia.

[o]AMV: avian myeloblastosis virus.

[p]STEMI: ST-elevation myocardial infarction.

[q]NSTEMI: non–ST-elevation myocardial infarction.

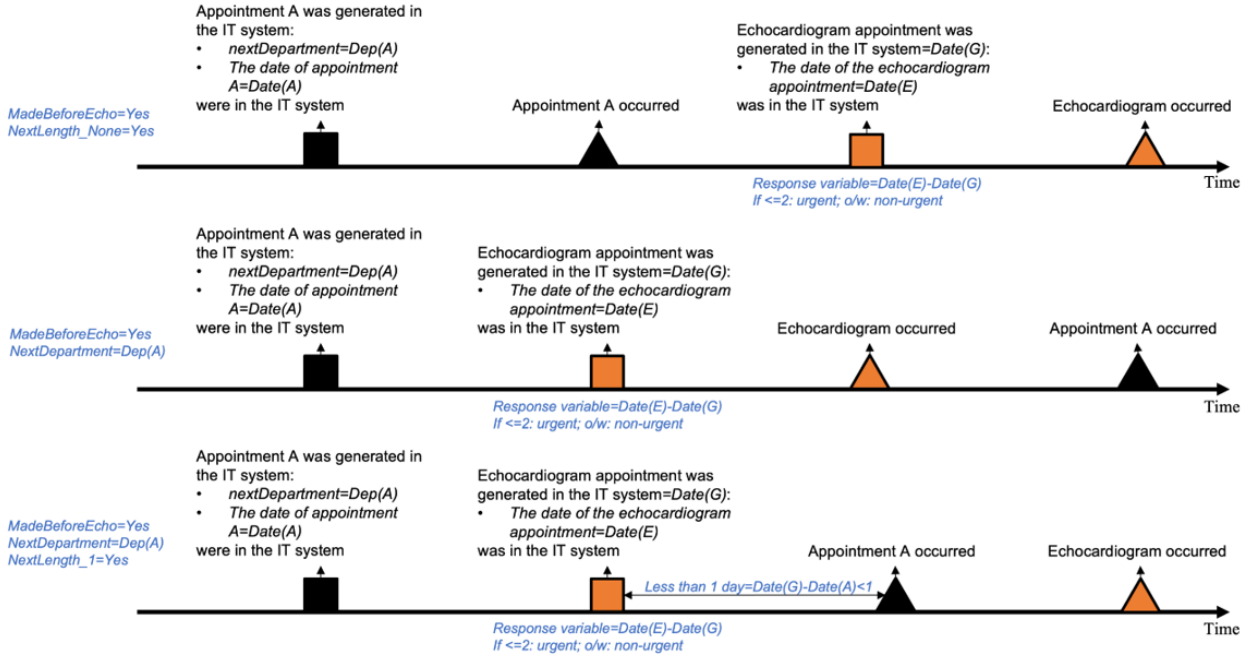
[r]PEA: pulseless electrical activity.

[s]AFib: atrial fibrillation.

[t]PDA: patent ductus arteriosus.

[u]SOB: shortness of breath.

Figure 1. Timeline and process of echocardiogram appointment scheduling. Using MadeBeforeEcho as an example.



Problem Formulation: Urgency Prediction Using OSDT

With data , where are M binary attributes and are the response variable, we model an OSDT tree d with a collection of H distinct leaves $d = (p_1,p_2,...,p_H)$. The objective function in this study integrates the misclassification error with a sparsity penalty imposed on the number of leaf nodes, denoted as $R(d,x,y)$. $R(d,x,y) = l(d,x,y) + \lambda H_d$, where $l(d,x,y)$ represents the misclassification error of the tree, which is computed as the fraction of training data with incorrectly predicted labels. In addition, H_d represents the number of leaves in tree d. To regularize the model and discourage larger trees, a regularization term λH_d is introduced, where λ is a hyperparameter controlling the strength of the penalty. A higher value of λ corresponds to a stronger penalty on the size of the tree. This implies that the tree is more likely to be shallower when achieving optimality.

By using OSDT, we aim to improve the overall performance of the classification task while simultaneously upholding a significant level of interpretability, thereby facilitating a comprehensive understanding of the underlying patterns and factors influencing the classification outcomes.

Results

Overview

In this section, we evaluated the proposed method against state-of-the-art machine learning models. We then highlighted attribute importance and provided clear interpretations of derived results within specific patient cohorts for transparency and clarity.

Performance Evaluation

We demonstrated the performance of our OSDT model by comparing it to commonly used machine learning models as baselines, including naive Bayes, generalized linear model, fast large margin, logistic regression, neural network, vanilla decision tree, random forest, gradient boosted trees, and support vector machine. The evaluation metrics used for the binary classification are accuracy, precision, recall, F_1-score, and F_2-score. Accuracy is a metric that quantifies the overall correctness of a machine learning model. It represented the proportion of correct predictions made by the model across all categories or classes. Precision and recall, on the other hand, measured the model's ability to accurately predict a specific category or class. Precision focused on the proportion of true positive predictions relative to all positive predictions made by the model. Recall, also known as sensitivity, gauged the model's capability to correctly detect instances of a specific category. It quantified the proportion of true positive predictions relative to all actual positive instances present in the data. The F_1-score has been widely used in the context of imbalanced classification problems and serves as a prominent metric. It is computed as the harmonic mean of the precision and recall scores, providing a balanced assessment of the model's performance by considering both precision and recall simultaneously. The F_2-score assigns greater weight to recall than precision, proving beneficial when the consequences of false negatives (ie, missed positive cases where patients are in urgent condition but remain unidentified by the model) outweigh those of false positives (ie, incorrectly identified positive cases). All metrics mentioned exhibited a range of values between 0 and 1, whereby a higher value indicated superior performance.

Compared with various baselines, the performance of the OSDT model achieved the highest accuracy, recall, F_1-score, and F_2-score (Table 2). The performance reported is based on 5-fold cross-validation. These results indicated the predictive capability of the OSDT model in our research context, demonstrating the overall performance and effectiveness of the OSDT model.

Table 2. OSDT[a] performance comparisons with baselines[b].

Algorithm	Accuracy (%), mean (SD)	Precision (%), mean (SD)	Recall (%), mean (SD)	F_1-score (%), mean (SD)	F_2-score[c] (%), mean (SD)
Naïve Bayes	78.86 (0.24)	81.3 (7.11)	3.34 (0.59)	6.41 (1.09)	4.13 (1.02)
Generalized linear model	79.23 (0.22)	78.05 (5.00)	5.93 (0.69)	11.01 (1.03)	7.27 (0.93)
Fast large margin	80.26 (0.47)	68.94 (2.57)	17.76 (1.4)	28.21 (1.7)	20.86 (2.17)
Logistic regression	79.26 (0.22)	77.68 (4.26)	6.16 (0.86)	11.41 (1.49)	7.55 (0.78)
Deep learning	80.49 (0.29)	85.59 (4.59)	12.14 (0.39)	21.26 (0.66)	14.66 (0.56)
Decision tree	80.69 (0.2)	69.18 (4.5)	22.45 (4.1)	33.53 (4.5)	25.96 (3.15)
Random forest	79.45 (0.18)	78.19 (5.54)	7.34 (0.31)	13.42 (0.57)	8.96 (2.67)
Gradient boosted trees	80.64 (0.29)	80.8 (2.96)	14.94 (1.55)	25.18 (2.25)	17.85 (1.95)
SVM[d]	80.3 (0.84)	61.42 (5.57)	24.06 (3.4)	34.48 (4.02)	27.39 (1.95)
OSDT (ours)	81.21 (0.20)	68.75 (1.7)	24.56 (0.59)	36.18 (0.66)	28.18 (0.55)

[a]OSDT: optimal sparse decision tree.

[b]OSDT is an algorithm that makes decisions based on direct constraints rather than generating probability scores. As a result, metrics like the receiver operating characteristic curve, precision and recall curve, and area under curve are not applicable for this method.Although the CIs for SVM and OSDT overlap, it is noteworthy that SVM exhibits a significantly larger SD. This indicates that OSDT is more robust in this scenario, delivering a more stable and reliable performance despite the overlapping intervals.

[c]; β=2.

[d]SVM: support vector machine.

Interpreting Prediction Results

OSDT, as a tree-based model, possesses the notable advantage of providing interpretable prediction results. We conducted an analysis of the decision trees generated using the entire dataset as well as specific patient cohorts. The objective is to extract the most influential rules that demonstrate both high accuracy and coverage, thereby aiming to uncover the underlying factors that drive the urgent decision of echocardiogram appointments.

We first identified several key categories and attributes that significantly influenced the urgency of patients' echocardiogram appointments (Table 3). First, the most important categories included "future scheduled process," pertaining to clinic scheduling policies, and "diagnosis," indicative of patients' health conditions. Second, within the top 12 important attributes, a cluster of attributes related to future scheduled processes emerged as the most prominent. These attributes encompassed scenarios if the next downstream appointment following the echocardiogram was scheduled prior to the echocardiogram appointment (ie, "MadeBeforeEcho"), instances where the next appointment did not pertain to the cardiovascular department (ie, "NextDepartment"), cases where no subsequent appointment was scheduled after the echocardiogram appointment (ie, "NextLength_None"), and situations where the time gap between the echo appointment and the subsequent one was less than a day ("NextLength_1"). The absence of a downstream appointment before the echocardiogram could be attributed to the clinic's practice of tailoring subsequent appointments based on the results of the echocardiogram. Consequently, it became imperative for medical providers to accord priority to the echocardiogram appointments of these patients, as the results would furnish vital evidence for guiding appropriate follow-up care and future steps. Third, attributes related to diagnoses assumed the second tier of importance, particularly whether patients exhibited respiratory and cardiac symptoms (ie, "R") or had documented cardiovascular conditions (ie, "I"). Patients diagnosed with heart-related issues, such as heart murmurs, shortness of breath, and chest pain, typically require expedited access to echocardiography results to determine the next course of action. Fourth, clinical setting attributes and demographic information are also important to patient prioritization. In the context of inpatients, health care providers tended to assign earlier echocardiogram appointment slots as part of a strategy to reduce the length of hospital stays. Additionally, when prioritizing patients with heart conditions, individuals referred by cardiologists received preferential treatment in terms of scheduling. Furthermore, the medical facility providing the data adopted a proactive approach by expediting echocardiogram appointments for out-of-state patients, aiming to minimize their duration of stay. This proactive stance facilitated timely evaluation and management, thereby contributing to a more efficient allocation of resources and an enhanced patient experience. Among medical history attributes, the presence of fluid and electrolyte disorders (ie, "FluidsLytes") emerged within the top 12, which underscored the strong correlation between fluid and electrolyte disorders and heart failure, further emphasizing its relevance in patient prioritization [29].

These results underscore the significance of admission and policy-related information in determining the urgency of echocardiogram appointments. They reflected the complexities of the scheduling process and highlighted the need for tailored appointment allocation strategies based on patients' referral status and downstream appointment requirements.

We subsequently focus on a specific patient cohort for further analysis. The "MadeBeforeEcho" attribute clearly emerged as

exceptionally significant among the dataset's attributes. It was noteworthy to highlight that, based on the data, there were no urgent cases when the "MadeBeforeEcho" variable was marked as "N." Consequently, we conducted an investigation specifically focusing on patients whose subsequent downstream appointment was scheduled before the date the echocardiogram appointment was generated in the system. This subset of the patient cohort served as an illustrative example of how decision trees could provide a high degree of interpretability in the context of patient prioritization (Figure 2). Upon scrutiny of the subdecision tree for this cohort depicted, several noteworthy observations emerged. Primarily, it became evident that the most crucial attribute for this cohort is "R," signifying whether the patient presents with respiratory and cardiac symptoms, which served as the root node of the subtree. The pathway leading to categorizing a patient case as urgent depended on multiple conditions: the patient exhibited respiratory and cardiac symptoms, had an appointment scheduled within the cardiology department, hailed from out of state, and had a subsequent appointment scheduled following the echocardiogram. In contrast, patients without respiratory and cardiac symptoms tended toward classification as nonurgent. This tendency toward nonurgency was particularly pronounced in cases lacking a scheduled appointment subsequent to the echocardiogram.

Table 3. Attribute importance and category importance[a].

Category and attribute	Meanings	Attribute importance
Future scheduled process (importance=0.0369)		
MadeBeforeEcho	Whether the next downstream appointment after echocardiogram is made before the date the echocardiogram appointment was generated in the system or not.	0.0279
NextDepartment	The department in which the appointment happened after the date the echocardiogram appointment was generated in the system.	0.0049
NextLength_None	No following appointment scheduled after the date the echocardiogram appointment was generated in the system.	0.0035
NextLength_1	The number of days from the date the echocardiogram appointment was generated in the system to its following appointment is less than 1 day.	0.0006
Diagnoses (importance=0.0154)		
R	If have murmur, hypoxemia, shortness, SOB[b], breath, shock, dyspnea, chest pain, troponin, syncope, electrocardiogram, extremity, mass, swelling, and edema.	0.0147
I	If have heart failure, coronary artery, cardiac arrest, STEMI[c], stroke, cardia, hypertension, endocarditis, NSTEMI[d], PEA[e] arrest, AFib[f], pulmonary embolism, pulmonary hypertension, and vegetation.	0.0007
Demographic (importance=0.0369)		
Geo_Out of State	Patient is from out of state.	0.0029
Geo_Town	Patient is from the local town.	0.0013
AGE_19-55	Age between 19 and 55 years.	0.0011
Clinical settings (importance=0.0053)		
ReferredType	Referred type-inpatient or outpatient.	0.0047
ReferredBy_CV	The specialty that patient referred by is cardiovascular disease department.	0.0006
FluidsLytes (medical history; importance=0.0021)	If have fluid and electrolyte disorders	0.0021

[a]The relative importance scores of the attribute category and individual attributes are determined by the Gini index of the optimal sparse decision tree. The feature importance values are relative importance values and do not have a fixed absolute range. We presented only the most important features.

[b]SOB: shortness of breath.

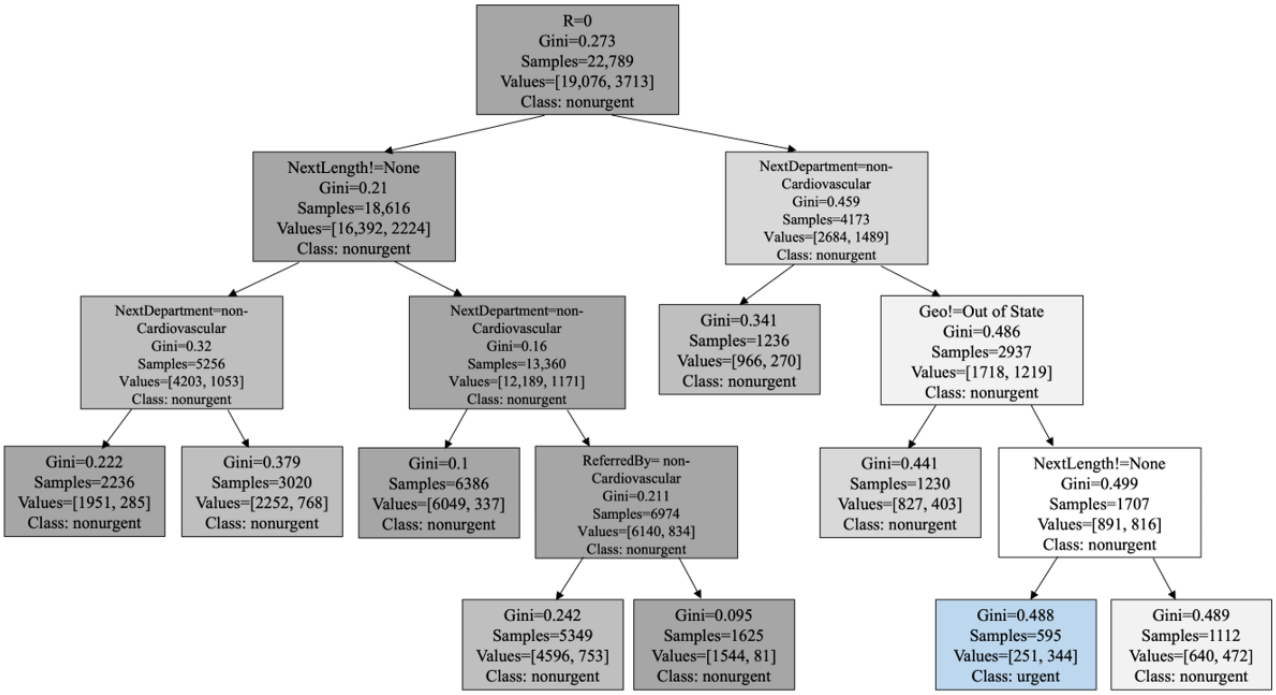
[c]STEMI: ST-elevation myocardial infarction.

[d]NSTEMI: non–ST-elevation myocardial infarction.

[e]PEA: pulseless electrical activity.

[f]AFib: atrial fibrillation.

Figure 2. The OSDT for patients whose next downstream appointment after the echocardiogram is scheduled before the date the echocardiogram appointment was generated in the system. OSDT: optimal sparse decision tree. λ=0.0008; accuracy: 83.69%.



Analyses on Diverse Patient Cohorts

In order to enhance the validity of the decision trees and gain more valuable medical insights, we conducted more analyses on smaller patient cohorts. Specifically, we focus on patients who have no next downstream appointment after echocardiogram and are categorized as inpatients. Furthermore, we narrowed down the patient cohort based on specific medical history and presented a compilation of rules extracted from the decision tree (Table 4).

A decision rule was defined as the pathway from the root of a decision tree to a leaf node . The accuracy and coverage of a decision rule served as critical metrics for evaluating its effectiveness and applicability. Accuracy, denoting the capacity of a decision rule to effectively forecast the outcome of interest, was quantified as the proportion of records that fulfill both the rule's precondition and its consequent within the precondition.

This metric was computed as , where "number of Correct Predictions" denoted the count of instances where the decision rule accurately anticipated the desired outcome and "Total number of Instances" represented the entire dataset or the set of instances under consideration, which elucidated how accuracy measures the precision of a decision rule in making predictions based on its specified conditions and its congruence with actual outcomes within the dataset. Coverage, on the other hand, measured the proportion of cases or individuals to which the decision rule could be applied. It could be calculated as . It signified the generalizability and practical scope of the rule in real-world scenarios. A decision rule with high coverage indicates its ability to be applied to a wide range of cases or individuals, thereby increasing its usefulness in practice.

In the context of patients with congestive heart failure (CHF), anemia played a significant role in determining the urgency of echocardiogram appointments (Table 4). Anemia could have detrimental effects on cardiac function through various mechanisms [29]. First, it induces cardiac stress by increasing heart rate and stroke volume. Additionally, anemia could lead to reduced renal blood flow and fluid retention, adding further strain to the heart. Prolonged anemia, regardless of its underlying cause, could contribute to the development of left ventricular hypertrophy, which exacerbates CHF by promoting cardiac cell death through apoptosis. Notably, patients with anemic CHF often exhibited resistance to CHF medications, and numerous studies consistently demonstrated that these individuals have a higher mortality rate compared to patients with non-anemic CHF [30]. Anemia also played a critical role in patients with coagulopathy, as it exacerbated bleeding, which in turn further worsens coagulopathy [30].

For patients with hypothyroidism, fluid and electrolyte disorders served as strong indicators. Hypothyroidism, a prevalent endocrine disorder, was associated with the development of congestive heart failure. Electrolyte disturbances were commonly observed in patients with chronic heart failure [31]. Echocardiogram has been a suitable modality for guiding fluid resuscitation in critically ill individuals. It allowed for the evaluation of fluid responsiveness based on several parameters, such as the left ventricle, aortic outflow, inferior vena cava, and right ventricle [32].

The impact of alcohol consumption on cardiovascular health was multifaceted. Extensive research has demonstrated that the consumption of alcohol at levels surpassing approximately 1 to 2 drinks per day was associated with hypertension [28]. This condition adversely affects the elasticity of arteries, leading to diminished blood and oxygen flow to the heart and consequently contributing to the onset of heart disease [33]. These pathophysiological changes increase the risk of heart disease. Consequently, patients with a history of alcohol abuse and

concomitant hypertension might require an urgent echocardiogram to assess the potential cardiac implications arising from these interconnected conditions.

Patients diagnosed with valvular heart conditions would fall into the urgent category if they also exhibited cardiovascular issues and a history of congestive heart failure. These attributes collectively signaled the presence of potentially serious cardiac problems, indicating a compelling need for an echocardiogram to obtain detailed cardiac information and facilitate accurate diagnoses. In the case of patients grappling with depression, their urgency classification as "urgent" was contingent upon the presence of co-occurring health issues. Extensive research has established a substantial influence of depression on the outcomes of concurrent medical conditions. Consequently, when depression coincided with other health problems, it necessitated an "urgent" classification, acknowledging its significant impact on overall health outcomes [34]. Regarding patients with obesity, an "urgent" classification applied if they additionally exhibited fluid and electrolyte disorders. Research findings have illuminated a connection between overweight or obesity and specific physiological factors, such as lower reactance and hypertonicity. Furthermore, individuals with overweight and those with obesity with lower reactance tended to demonstrate significantly elevated serum sodium levels compared to individuals with a normal weight. These associations underscored the importance of promptly addressing the medical needs of patients with obesity with fluid and electrolyte disorders, warranting an "urgent" classification for their cases [35].

Overall, the decision rules extracted from our analyses aligned closely with medical knowledge, providing reliable insights for identifying urgent echocardiogram appointments for patients. The congruence between the rules and medical understanding not only validated the effectiveness of our model but also highlighted the consistent application of medical principles in the decision-making process. This focused analysis contributed to a better understanding of the OSDT model's validity and offered valuable medical perspectives to enhance the identification of urgent patients' echocardiogram appointments.

Table 4. Decision rules for specific patient cohorts.

Cohort	Rules for a patient to be classified as urgent	Rule accuracy (%)	Rule coverage (%)
CHF[a]	The department in which the appointment happened after the echocardiogram appointment was generated in the system=non-cardiovascular disease, AGE<75, anemia=yes	100	14.20
Coagulopathy	Anemia=Yes	99	53.03
Hypothyroid	Fluid and electrolyte disorders=yes, Whether the patient had a cardiovascular surgery within six months prior to the echocardiogram appointment=no	100	32.91
Alcohol	Hypertension=yes	100	43.75
Valvular	I=1(has cardiovascular conditions), CHF=yes	100	6.36
Depression	Z=1 (has factors influencing health status and contact with health service)	100	24.49
Obesity	Geo!=Town, E=0 (has no nutritional and metabolic diseases), fluid and electrolyte disorders=yes	100	23.75

[a]CHF: congestive heart failure.

Discussion

Overview

The primary objective of our study is to forge an effective tree-based classification machine learning model geared toward prioritizing the allocation of echocardiogram appointments for patients with a heightened need for timely diagnostics. Our long-term goal is to streamline the scheduling process, ensuring that patients' medical requirements are promptly addressed, thereby minimizing delays and optimizing their health care experience. Moreover, our study aspired to delve deeper into the intricate attributes that contribute to the urgency of echocardiogram lab appointments. Recognizing the intricate interplay of medical, logistical, and patient-specific variables, we sought to unravel the complex rules and dynamics that govern appointment prioritization. By harnessing the inherent interpretability of our model, we aim to uncover hidden insights and relationships within a large amount of EHR data, shedding light on the critical determinants that underscore the need for rapid scheduling. The implications of our study extended beyond the realm of predictive modeling. We aimed to empower health care professionals with a powerful tool that not only optimizes resource allocation but also enriches their decision-making process.

Principal Results

The findings demonstrate promising results by accurately predicting the urgency of echocardiogram appointments and providing valuable insights into the critical guidelines applicable to specific patient cohorts. In summary, the study emphasizes two key points: (1) among the various attributes examined, it is observed that admission-related attributes exert a significant influence on the level of urgency for patients' echocardiogram appointments; and (2) the urgency of scheduling echocardiogram appointments can be influenced by the presence of comorbidities that exacerbate patients' conditions. In the case of congestive heart failure, anemia emerges as a significant attribute, highlighting its relevance in contributing to the urgency of echocardiogram appointments. Similarly, coagulopathy is identified as an important attribute for patients with congestive heart failure, further emphasizing the need for prompt

assessment. For patients with hypothyroidism, the presence of fluid and electrolyte disorders serves as a concerning indicator, warranting the prioritization of an echocardiogram. Additionally, hypertension is found to be a critical medical knowledge for patients with a history of alcohol abuse, underscoring the urgency of echocardiogram in this population.

Our work is unique in applying an advanced binary decision tree model that offers inherent interpretability, avoiding the limitations of post hoc techniques like local interpretable model-agnostic Explanation and Shapley additive explanation, such as local interpretability constraints, sensitivity to perturbations, and difficulties in selecting appropriate surrogate models. We extract interpretable rules grounded in medical knowledge, making this the first study to introduce tree-based interpretable machine learning for patient prioritization and the stratification of medical test urgency. Furthermore, the tree-based model allows us to derive rules that are easily understandable to medical professionals. These rules can be assessed for alignment with existing medical knowledge and applied in real-world practice by health care providers.

Limitations

The research has several limitations that could be addressed in future work. First, the accuracy of the prediction model hinges on the quality and completeness of available data; incomplete or missing data may compromise the reliability of predictions. Furthermore, it is essential to recognize that the effectiveness of the model may vary when applied to diverse patient populations or health care settings. This variation can be attributed to the unique attributes and patterns present in the training data, which significantly impact the model's performance. Moreover, the predictions rely on the elapsed days between the appointment scheduling date and the appointment date. Nonurgent patients may inadvertently be grouped with urgent cases due to cancellations and rescheduling of echocardiogram appointments. While this offers a broad indication of urgency, it may overlook critical factors that influence appointment priority. Integrating essential clinical or contextual details, such as the patient's medical history, symptom severity, or health care resource availability, into the model could provide more comprehensive insights.

Conclusions

This research adapts the OSDT algorithm to assess the urgency of patients in need of echocardiograms. The OSDT model demonstrates better performance over alternative machine learning models, highlighting its predictive accuracy and effectiveness. Furthermore, it identifies key attributes and rules governing the prioritization of echocardiogram appointments.

The analysis of decision trees generated by the OSDT model reveals the significance of admission- and policy-related attributes, such as downstream appointment scheduling and patient referral status, in determining appointment urgency. Moreover, the analyses of specific patient cohorts provide medical insights into the role of comorbidities, such as anemia in patients with CHF and coagulopathy, and fluid and electrolyte disorders in patients with hypothyroidism. These insights align with established medical knowledge and enhance the identification of urgent echocardiogram appointments.

In summary, this study facilitates the development of effective scheduling protocols for echocardiogram appointments by harnessing machine learning techniques and integrating medical insights. This approach enhances the overall efficiency and effectiveness of echocardiogram services, ultimately benefiting patient care. The findings can also be generalized to inform the establishment of efficient scheduling protocols and the promotion of equitable access to various other medical laboratory tests.

Acknowledgments

In this research, the authors gratefully acknowledge the financial support provided by the Ivy College of Business and the Department of Industrial and Manufacturing Systems Engineering at Iowa State University. The authors also extend our appreciation to the Mayo Clinic for generously providing essential data. Their collaborative efforts significantly enriched our study.

Conflicts of Interest

None declared.

References

1. Danzon PM, Manning WG, Marquis MS. Factors affecting laboratory test use and prices. Health Care Financ Rev 1984;5(4):23-32 [FREE Full text] [Medline: 10317549]
2. Bhatt J, Bathija P. Ensuring access to quality health care in vulnerable communities. Acad Med 2018;93(9):1271-1275 [FREE Full text] [doi: 10.1097/ACM.0000000000002254] [Medline: 29697433]
3. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. Artif Intell Healthc Elsevier 2020:25-60. [doi: 10.1016/b978-0-12-818438-7.00002-2]
4. Ashley EA, Niebauer J. Cardiology Explained. London, United Kingdom: Remedica; 2004.
5. Cheitlin MD, Armstrong WF, Aurigemma GP, Beller GA, Bierman FZ, Davis JL, et al. ACC/AHA/ASE 2003 guideline update for the clinical application of echocardiography: summary article. A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (ACC/AHA/ASE Committee to update the 1997 guidelines for the clinical application of echocardiography). J Am Soc Echocardiogr 2003;16(10):1091-1110. [doi: 10.1016/S0894-7317(03)00685-0] [Medline: 14566308]

6. Aluru JS, Barsouk A, Saginala K, Rawla P, Barsouk A. Valvular heart disease epidemiology. Med Sci 2022;10(2):32 [FREE Full text] [doi: 10.3390/medsci10020032] [Medline: 35736352]
7. Pushparajah K, Garvie D, Hickey A, Qureshi SA. Managed care network for the assessment of cardiac problems in children in a district general hospital: a working model. Arch Dis Child 2006;91(11):892-895 [FREE Full text] [doi: 10.1136/adc.2005.086058] [Medline: 16717084]
8. Murugan SJ, Thomson J, Parsons JM, Dickinson DF, Blackburn MEC, Gibbs JL. New outpatient referrals to a tertiary paediatric cardiac centre: evidence of increasing workload and evolving patterns of referral. Cardiol Young 2005;15(1):43-46. [doi: 10.1017/S1047951105000090] [Medline: 15831160]
9. Mariotti G, Siciliani L, Rebba V, Fellini R, Gentilini M, Benea G, et al. Waiting time prioritisation for specialist services in Italy: the homogeneous waiting time groups approach. Health Policy 2014;117(1):54-63. [doi: 10.1016/j.healthpol.2014.01.018] [Medline: 24576498]
10. Solans-Domènech M, Adam P, Tebé C, Espallargues M. Developing a universal tool for the prioritization of patients waiting for elective surgery. Health Policy 2013;113(1-2):118-126. [doi: 10.1016/j.healthpol.2013.07.006] [Medline: 23932414]
11. Silva-Aravena F, Morales J. Dynamic surgical waiting list methodology: a networking approach. Mathematics 2022;10(13):2307. [doi: 10.3390/math10132307]
12. Holt JM, Wilk B, Birch CL, Brown DM, Gajapathy M, Moss AC, Undiagnosed Diseases Network, et al. VarSight: prioritizing clinically reported variants with binary classification algorithms. BMC Bioinformatics 2019;20(1):496 [FREE Full text] [doi: 10.1186/s12859-019-3026-8] [Medline: 31615419]
13. Abbasgholizadeh Rahimi S, Jamshidi A, Ruiz A, Ait-kadi D. A new dynamic integrated framework for surgical patients' prioritization considering risks and uncertainties. Decis Support Syst 2016;88:112-120. [doi: 10.1016/j.dss.2016.06.003]
14. Rabbani N, Kim GYE, Suarez CJ, Chen JH. Applications of machine learning in routine laboratory medicine: current state and future directions. Clin Biochem 2022;103:1-7 [FREE Full text] [doi: 10.1016/j.clinbiochem.2022.02.011] [Medline: 35227670]
15. Javaid M, Haleem A, Pratap Singh R, Suman R, Rab S. Significance of machine learning in healthcare: features, pillars and applications. Int J Intell Netw 2022;3:58-73. [doi: 10.1016/j.ijin.2022.05.002]
16. Elitzur R, Krass D, Zimlichman E. Machine learning for optimal test admission in the presence of resource constraints. Health Care Manag Sci 2023;26(2):279-300 [FREE Full text] [doi: 10.1007/s10729-022-09624-1] [Medline: 36631694]
17. Marescotti D, Narayanamoorthy C, Bonjour F, Kuwae K, Graber L, Calvino-Martin F, et al. AI-driven laboratory workflows enable operation in the age of social distancing. SLAS Technol 2022;27(3):195-203 [FREE Full text] [doi: 10.1016/j.slast.2021.12.001] [Medline: 35058197]
18. Zhang K, Jiang X, Madadi M, Chen L, Savitz S, Shams S. DBNet: a novel deep learning framework for mechanical ventilation prediction using electronic health records. 2021 Presented at: BCB '21: 12th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; August 1-4, 2021; Gainesville, FL p. 1-8. [doi: 10.1145/3459930.3469551]
19. Azimi V, Zaydman M. Optimizing equity: working towards fair machine learning algorithms in laboratory medicine. J Appl Lab Med 2023;8(1):113-128. [doi: 10.1093/jalm/jfac085] [Medline: 36610413]
20. Chaddad A, Peng J, Xu J, Bouridane A. Survey of explainable AI techniques in healthcare. Sensors 2023;23(2):634 [FREE Full text] [doi: 10.3390/s23020634] [Medline: 36679430]
21. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. 2020 Presented at: AIES '20: AAAI/ACM Conference on AI, Ethics, and Society; February 7-9, 2020; New York, NY p. 180-186. [doi: 10.1145/3375627.3375830]
22. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1(5):206-215 [FREE Full text] [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]
23. Wiedermann W, Bonifay W, Huang FL. Advanced categorical data analysis in prevention science. Prev Sci 2023;24(3):393-397. [doi: 10.1007/s11121-022-01485-y] [Medline: 36633766]
24. Hu X, Rudin C, Seltzer M. Optimal sparse decision trees. ArXiv Preprint posted online on October 1, 2019. [doi: 10.5260/chara.21.2.8]
25. Dosilovic FK, Brcic M, Hlupic N. Explainable artificial intelligence: a survey. 2009 Presented at: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO); May 21-25, 2018; Opatija, Croatia p. 0210-0215.
26. Podgorelec V, Kokol P, Stiglic B, Rozman I. Decision trees: an overview and their use in medicine. J Med Syst 2002;26(5):445-463. [doi: 10.1023/a:1016409317640] [Medline: 12182209]
27. Lavanya D, Rani KU. Performance evaluation of decision tree classifiers on medical datasets. IJCA 2011;26(4):1-4. [doi: 10.5120/3095-4247]
28. Piano MR. Alcohol's effects on the cardiovascular system. Alcohol Res 2017;38(2):219-241 [FREE Full text] [Medline: 28988575]
29. Urso C, Brucculeri S, Caimi G. Acid-base and electrolyte abnormalities in heart failure: pathophysiology and implications. Heart Fail Rev 2015;20(4):493-503 [FREE Full text] [doi: 10.1007/s10741-015-9482-y] [Medline: 25820346]

30. Silverberg D, Wexler D, Iaina A, Schwartz D. The role of anemia in the progression of congestive heart failure: Is there a place for erythropoietin and intravenous iron? Transfus Altern Transfus Med 2008;6(3):26-37. [doi: 10.1111/j.1778-428x.2005.tb00121.x]
31. Costache II, Cimpoeşu D, Petriş O, Petriş AO. Electrolyte disturbances in patients with chronic heart failure—clinical, evolutive and therapeutic implications. Rev Med Chir Soc Med Nat Iasi 2012;116(3):708-713. [Medline: 23272514]
32. Miller A, Mandeville J. Predicting and measuring fluid responsiveness with echocardiography. Echo Res Pract 2016;3(2):G1-G12 [FREE Full text] [doi: 10.1530/ERP-16-0008] [Medline: 27249550]
33. Petrie JR, Guzik TJ, Touyz RM. Diabetes, hypertension, and cardiovascular disease: clinical insights and vascular mechanisms. Can J Cardiol 2018;34(5):575-584 [FREE Full text] [doi: 10.1016/j.cjca.2017.12.005] [Medline: 29459239]
34. Cassano P, Fava M. Depression and public health: an overview. J Psychosom Res 2002;53(4):849-857. [doi: 10.1016/s0022-3999(02)00304-5] [Medline: 12377293]
35. Stookey JD, Barclay D, Arieff A, Popkin BM. The altered fluid distribution in obesity may reflect plasma hypertonicity. Eur J Clin Nutr 2007;61(2):190-199. [doi: 10.1038/sj.ejcn.1602521] [Medline: 17021599]

Abbreviations

CHF: congestive heart failure
EHR: electronic health record
OSDT: optimal sparse decision tree

Edited by Z Yin; submitted 10.07.24; peer-reviewed by Y Li, M Madadi; comments to author 05.09.24; revised version received 18.10.24; accepted 16.12.24; published 29.01.25.

Please cite as:
Jiang Y, Li Q, Huang YL, Zhang W
Urgency Prediction for Medical Laboratory Tests Through Optimal Sparse Decision Tree: Case Study With Echocardiograms
JMIR AI 2025;4:e64188
URL: https://ai.jmir.org/2025/1/e64188
doi:10.2196/64188
PMID:39879091

Original Paper

Advancing Privacy-Preserving Health Care Analytics and Implementation of the Personal Health Train: Federated Deep Learning Study

Ananya Choudhury[1,2*], MTech; Leroy Volmer[1,2*], MSc; Frank Martin[3], MSc; Rianne Fijten[1,2], PhD; Leonard Wee[1,2], PhD; Andre Dekker[1,2,4], PhD; Johan van Soest[1,2,4], PhD

[1]GROW Research Institute for Oncology and Reproduction, Maastricht University Medical Center+, Maastricht, Netherlands

[2]Clinical Data Science, Maastricht University, Maastricht, Netherlands

[3]Netherlands Comprehensive Cancer Organization (IKNL), Eindhoven, Netherlands

[4]Brightlands Institute for Smart Society (BISS), Faculty of Science and Engineering (FSE), Maastricht University, Heerlen, Netherlands

[*]these authors contributed equally

Corresponding Author:
Ananya Choudhury, MTech
GROW Research Institute for Oncology and Reproduction
Maastricht University Medical Center+
Paul Henri Spakalaan 1
Maastricht, 6229EN
Netherlands
Phone: 31 0686008485
Email: ananya.aus@gmail.com

Abstract

Background: The rapid advancement of deep learning in health care presents significant opportunities for automating complex medical tasks and improving clinical workflows. However, widespread adoption is impeded by data privacy concerns and the necessity for large, diverse datasets across multiple institutions. Federated learning (FL) has emerged as a viable solution, enabling collaborative artificial intelligence model development without sharing individual patient data. To effectively implement FL in health care, robust and secure infrastructures are essential. Developing such federated deep learning frameworks is crucial to harnessing the full potential of artificial intelligence while ensuring patient data privacy and regulatory compliance.

Objective: The objective is to introduce an innovative FL infrastructure called the Personal Health Train (PHT) that includes the procedural, technical, and governance components needed to implement FL on real-world health care data, including training deep learning neural networks. The study aims to apply this federated deep learning infrastructure to the use case of gross tumor volume segmentation on chest computed tomography images of patients with lung cancer and present the results from a proof-of-concept experiment.

Methods: The PHT framework addresses the challenges of data privacy when sharing data, by keeping data close to the source and instead bringing the analysis to the data. Technologically, PHT requires 3 interdependent components: “tracks” (protected communication channels), “trains” (containerized software apps), and “stations” (institutional data repositories), which are supported by the open source “Vantage6” software. The study applies this federated deep learning infrastructure to the use case of gross tumor volume segmentation on chest computed tomography images of patients with lung cancer, with the introduction of an additional component called the secure aggregation server, where the model averaging is done in a trusted and inaccessible environment.

Results: We demonstrated the feasibility of executing deep learning algorithms in a federated manner using PHT and presented the results from a proof-of-concept study. The infrastructure linked 12 hospitals across 8 nations, covering 4 continents, demonstrating the scalability and global reach of the proposed approach. During the execution and training of the deep learning algorithm, no data were shared outside the hospital.

Conclusions: The findings of the proof-of-concept study, as well as the implications and limitations of the infrastructure and the results, are discussed. The application of federated deep learning to unstructured medical imaging data, facilitated by the PHT framework and Vantage6 platform, represents a significant advancement in the field. The proposed infrastructure addresses the

challenges of data privacy and enables collaborative model development, paving the way for the widespread adoption of deep learning–based tools in the medical domain and beyond. The introduction of the secure aggregation server implied that data leakage problems in FL can be prevented by careful design decisions of the infrastructure.

Trial Registration: ClinicalTrials.gov NCT05775068; https://clinicaltrials.gov/study/NCT05775068

(JMIR AI 2025;4:e60847) doi:10.2196/60847

KEYWORDS

gross tumor volume segmentation; federated learning infrastructure; privacy-preserving technology; cancer; deep learning; artificial intelligence; lung cancer; oncology; radiotherapy; imaging; data protection; data privacy

Introduction

Federated learning (FL) allows the collaborative development of artificial intelligence models using large datasets, without the need to share individual patient-level data [1-4]. In FL, partial models trained on separate datasets are shared, but not the data itself, hence a global model is derived from the collective set of partial models. This study introduces an innovative FL framework known as the Personal Health Train (PHT) that includes the procedural, technical, and governance components needed to implement FL on real-world health care data, including the training of deep learning neural networks [5]. The PHT infrastructure is supported by a free and open-source infrastructure known as "priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange," that is, Vantage6 [6]. We will describe in detail an architecture for training a deep learning model in a federated way with 12 institutional partners located in different parts of the world.

Sharing patient data between health care institutions is tightly regulated due to concerns about patient confidentiality and the potential for misuse of data. Data protection laws—including the European Union's General Data Protection Regulations; Health Insurance Portability and Accountability Act of 1996 (HIPAA) in the United States; and similar regulations in China, India, Brazil, and many other countries—place strict conditions on the sharing and secondary use of patient data [7]. Incompatibilities between laws and variations in the interpretation of such laws lead to strong reluctance about sharing data across organizational and jurisdictional boundaries [8-10].

To address the challenges of data privacy, a range of approaches have been published in the literature. Differential privacy, homomorphic encryption, and FL comprise a family of applications known as "privacy enhancing technologies" [11-13]. The common goal of privacy-enhancing technologies is to unlock positively impactful societal, economic, and clinical knowledge by analyzing data en masse, while obscuring the identity of study subjects that make up the dataset. Academic institutions are more frequently setting up controlled workspaces (eg, secure research environments [SREs]), where multiple researchers can collaborate on data analysis within a common cloud computing environment, but without allowing access to the data from outside the SRE desktop; however, this assumes that all the data needed have been transferred into the SRE in the first place [14,15]. Similarly, the National Institutes of Health has set up an "Imaging Data Commons" to provide secure access to a large collection of publicly available cancer imaging data colocated with analysis tools and resources [16]. Other researchers have shown that blockchain encryption technology can be used to securely store and share sensitive medical data [17]. Blockchain ensures data integrity by maintaining an audit trail of every transaction, while zero trust principles make sure the medical data are encrypted and only authenticated users and devices interact with the network [18].

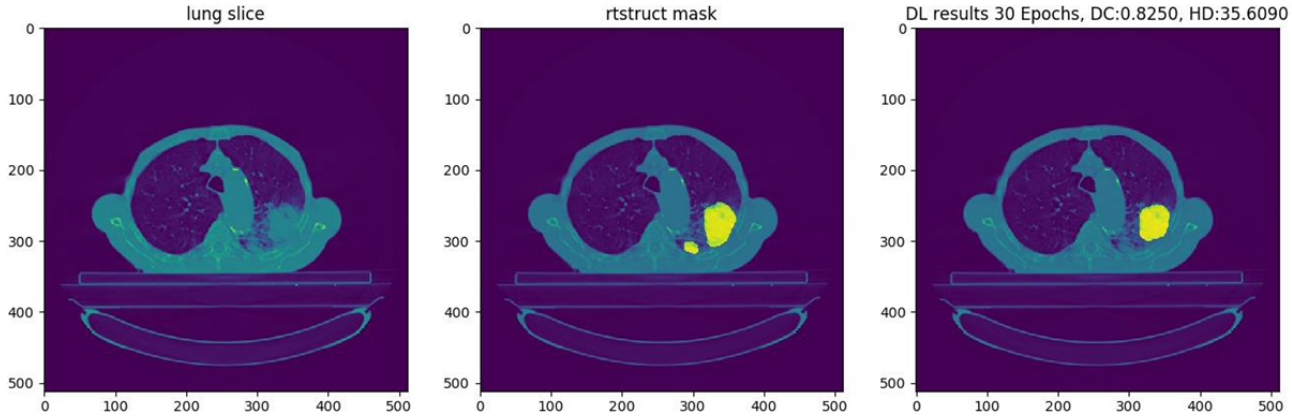
From a procedural point of view, the PHT manifesto for FL rules out the sharing of individual patient-level data between institutions, no matter if the patient data have been deidentified or encrypted [19]. The privacy-by-design principle here may be referred to as "safety in numbers," that is, any single individual's data values are obscured, by computing either the descriptive statistics or the partial model, over multiple patients. PHT allows sufficiently adaptable methods of model training, such as iterative numerical approximation (eg, bisection) or federated averaging (FedAvg [20]), and does not mandatorily require model gradients or model residuals, which are well-known avenues of privacy attacks [21-24]. Governance is essential with regards to compliance with privacy legislation and division of intellectual property between collaboration partners. A consortium agreement template for PHT has been made openly accessible [25], which is based on our current consortium ARGOS (artificial intelligence for gross tumor volume segmentation) [26]. Technologically, PHT requires 3 interdependent components to be installed—"tracks" are protected telecommunications channels that connect partner institutions, "trains" are Docker containerized software apps that execute a statistical analysis that all partners have agreed upon, and "stations" are the institutional data repositories that hold the patient data [23]. It is this technological infrastructure—the tracks, trains, and stations—that is supported by the aforementioned Vantage6 software, for which detailed stand-alone documentation exists [27].

The paper proposes a federated deep learning infrastructure based on the PHT manifesto [19], which provides a governance and ethical, legal, and social implications framework for conducting FL studies across geographically diverse data providers. The research aims to showcase a custom FL infrastructure using the open-source Vantage6 platform, detailing its technological foundations and implementation specifics. The paper emphasizes the significance of the implemented custom federation strategy, which maintains a strict separation between intermediate models from both internal and external user access. This approach is crucial for safeguarding the security and privacy of sensitive patient data,

as it prevents potential reverse engineering of intermediate results that could compromise confidentiality. This aggregation strategy is particularly important in the case of deep learning–based studies where multiple iterations of models or gradients are necessary to derive an optimal global model.

To demonstrate the infrastructure's robustness and practical applicability, the study presents a proof-of-concept involving the development of a federated deep learning algorithm based on 2D convolutional neural network (CNN) architecture [28]. This algorithm was implemented to automatically segment gross tumor volume (GTV) from lung computed tomography (CT) images of patients with lung cancer. Figure 1 [29] demonstrates a manual segmentation and deep learning–based segmentation of a tumor in the chest CT image of a patient. The subsequent sections provide a comprehensive account of the precise technical specifications of the infrastructure that links 12 hospitals across 8 nations, covering 5 continents. The algorithm developed learns from the distributed datasets and deploys it using the infrastructure. However, it is important to mention that the choice of the use case is only exemplary in nature, and the infrastructure is equipped to train any kind of deep learning architecture for relevant clinical use cases.

Figure 1. Illustrative result on a hold-out validation slice; the main bulk of the gross tumor volume as determined by the oncologist (middle) has been correctly delineated by the deep learning algorithm (right), but a small tumor mass adjacent and to the lower right of the main gross tumor volume mass has been missed (reproduced from Figure 6 of Chapter 4 of the thesis by Patil [29], which is published under the Taverne License [Article 25fa of the Dutch Copyright Act]).



The research used a deep learning architecture because in recent times the application of deep learning in health care has led to impressive results, specifically in the areas of natural language processing and computer vision (medical image analysis), with the promise for more efficient diagnostics and better predictions of treatment outcomes in future [30-35]. However, for robust generalizability, and to earn clinicians' acceptance, it is essential that artificial intelligence apps are trained on massive volumes of diverse and demographically representative health care data across multiple institutions. Given the barriers to data sharing, this is clearly an area where FL can play a vital role. Many studies have been published that present FL on medical data including federated deep learning [36-40]. However, only a limited number of studies have documented the use of dedicated frameworks and infrastructures in a transparent manner. The adoption of a custom federation strategy or absence of explicit reporting on the used infrastructure is observed in most of the studies. Table 1 summarizes the small number of FL studies that have been published in connection with deep learning investigations related to medical image segmentations to date.

The paper primarily focuses on demonstrating the training and aggregation mechanism of a deep learning architecture within a FL framework. It deliberately avoids delving into the optimization of model performance or clinical accuracy, as these aspects fall outside the paper's scope. Instead of emphasizing the selection of an optimal CNN architecture or aggregation strategy [39], the research concentrates on elucidating the functionality of the FL infrastructure. Existing literature has shown that FL models can achieve performance comparable to centrally trained models [38,41,45-47]. This supports the assumption that, given identical datasets and CNN architectures, a model trained using FL would likely yield similar results to one trained through centralized methods. The paper operates under this premise, prioritizing the explanation of the FL process over demonstrating performance parity with centralized training approaches.

The study highlights 3 key points as follows:

- FL is particularly well suited for deep learning applications, which typically require vast amounts of data. This makes it an ideal showcase for the federated approach.
- When implementing federated deep learning, it is crucial to have a robust infrastructure and use a customized, secure aggregation strategy. These elements are essential for safeguarding the privacy of sensitive patient information.
- FL in real-world medical data is not just a technological challenge; it requires a comprehensive strategy that addresses ethical, legal, governance, and organizational aspects, as highlighted by the PHT manifesto.

Table 1. Existing studies from the literature focusing on federated deep learning on medical images.

Infrastructure and clinical use case	Data type	Scale
NVIDIA FLARE/CLARA		
Prostate segmentation of T2-weighted MRI[a] [41]	DICOM MRI	3 centers
COVID-19 pneumonia detection [42]	Chest CT[b]	7 centers
Tensorflow federated		
COVID-19 prediction from chest CT images [43]	Chest CT	3 datasets
OpenFL		
Glioblastoma tumor boundary detection [44]	Brain MRI	71 centers

[a]MRI: magnetic resonance imaging.

[b]CT: computed tomography.

The findings of the proof-of-concept study, as well as the implications and limitations of the infrastructure and the results, are discussed. The subsequent section of the paper is structured as follows: the *Methods* section describes the approach taken, followed by the *Results*, which detail the implementation of the infrastructure and a proof-of-concept execution. Finally, the paper concludes with a *Discussion* section.

Methods

Overview

When conducting a federated deep learning study, it is crucial to consider several key perspectives, which include both technical as well as organizational and legal aspects. These key factors have been instrumental in designing the infrastructure architecture used for training the deep learning algorithm. In this section, we discuss the technical details while adhering to an Ethics-Legal-Social Impact framework as laid down by the PHT manifesto. The technical design decisions are based on the following assumptions:

Data Landscape

Understanding the data landscape is crucial in designing and deploying FL algorithms. The technological approaches for handling horizontally partitioned data, where each institution contains nonoverlapping human subjects but the domain of the data (eg, CT images of lung cancer) is the same across different institutions, can differ significantly from those used for vertically partitioned data, where each institution contains the same human subjects but the domain of the data do not overlap (eg, CT scans in one, but socioeconomic metrics in another). Additionally, unstructured data, such as medical images, requires different algorithms and preprocessing techniques compared with structured data. In this paper, the architecture will only focus on CT scans and horizontally partitioned patient data.

Data Preprocessing

In a horizontally partitioned FL setting, the key preprocessing steps can be standardized and sent to all partner institutions. However, the workflow needs to handle differences in patients, scan settings, and orientations. Anonymization, quality improvements, and DICOM standardization ensure homogeneity and high quality across hospitals. These offline preprocessing steps, applied consistently to the horizontally partitioned data, enabled using the same model across institutions, crucial for the FL study's success.

Network Topology of the FL Infrastructure

The network topology choice for implementing FL can vary from client-server, peer-to-peer, tree-based hierarchical, or hybrid topologies. While peer-to-peer architecture is more cost-effective and offers a high capacity, it has the disadvantages of a lack of security and privacy constraints and a complex troubleshooting process in the event of a failure. The choice of network topology for this study is based on a client-server architecture, offering a single point of control in the form of the central server.

Choice of Model Aggregation Site

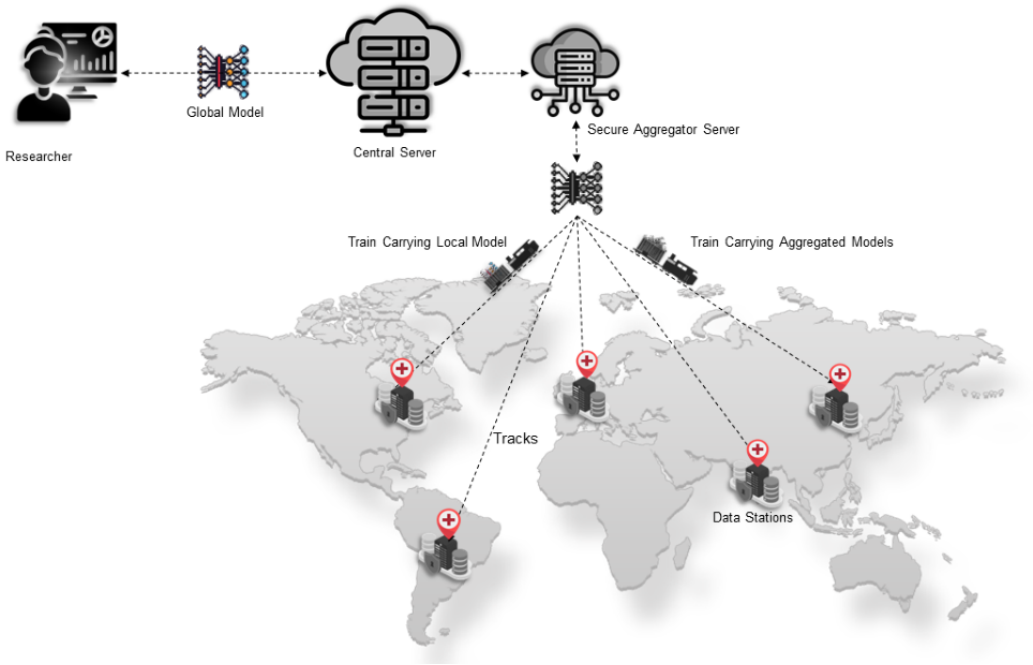
For a client-server architecture, the model aggregation can occur either in one of the data providers' machines, the central server, or in a dedicated aggregation server. For this implementation, we opted to use a dedicated aggregation server. The details and benefits of the implementation are discussed in the next section.

Training Strategy

The communication mechanism for transferring weights can be either synchronous, asynchronous, or semisynchronous, and weights can be consolidated using ensemble learning, FedAvg, split learning, weight transfer, or swarm learning. The strategy used for this study is based on a synchronous mechanism using the FedAvg algorithm. This gives a simple approach, where the averaging algorithm waits for all the data centers to transfer the locally trained model before initiating the averaging.

Based on the assumption, Figure 2 depicts the overall architecture of the federated deep learning study presented in the paper. The next section describes the FL Infrastructure in detail.

Figure 2. Overall architecture of ARGOS (artificial intelligence for gross tumor volume segmentation) federated deep learning architecture adapted from Vantage6. The figure depicts a researcher connected to the central server, a secure aggregation server, trains carrying models, connected data stations, and the communicating tracks.



The ARGOS Federated Deep Learning Infrastructure

Overview

In accordance with the PHT principles, the ARGOS infrastructure is comprised of 3 primary categories of components, labeled as the data stations, the trains, and the track. Furthermore, the architectural framework encompasses various roles that map to the level of permissions and access, specifically a track provider, the data providers, and the researcher. The infrastructure implementation can be further categorized into 3 important components: a central coordination server, a secure aggregation server (SAS), and the nodes located at each "data station." In the following sections, we attempt to describe each of these components and the respective stakeholders responsible for maintaining them.

Central Coordinating Server

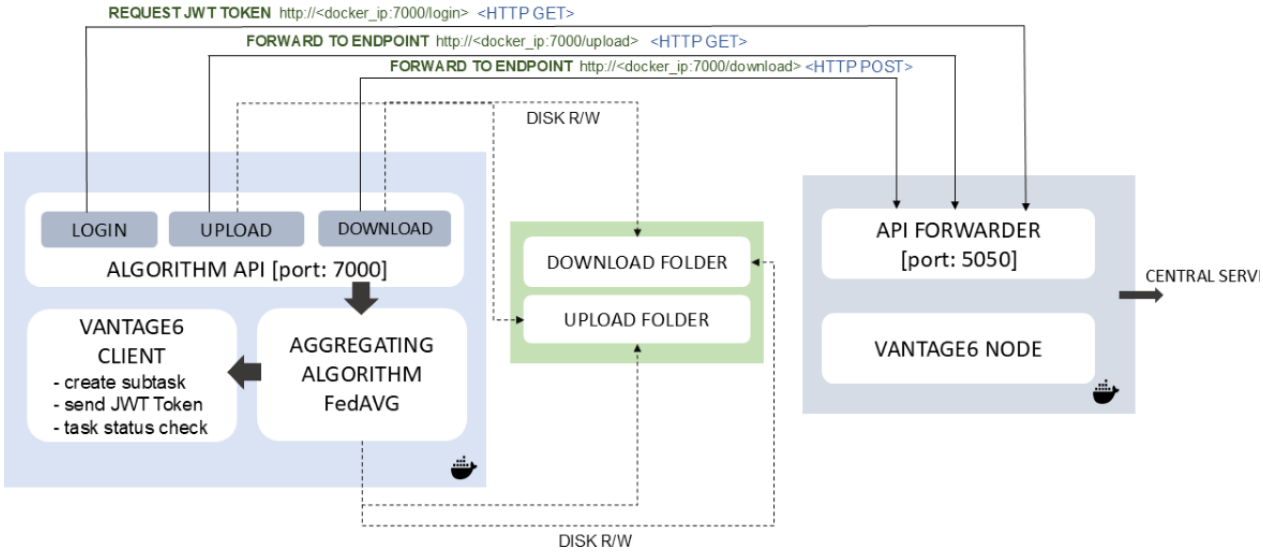
The central coordination server is located at the highest hierarchical level and serves as an intermediary for message exchange among all other components. The components of the system, including the users, data stations, and SAS, are registered entities that possess well-defined authentication mechanisms within the central server. It is noteworthy that the central acts as a coordinator rather than a computational engine. Its primary function is to store task-specific metadata relevant to the task initiated for training the deep learning algorithm. In the original Vantage6 infrastructure, the central server also stores the intermediate results. In the ARGOS infrastructure, the central server is designed to not store any intermediate results but only the global aggregated model at the end of the entire training process.

Secure Aggregation Server

The SAS refers to a specialized station that contains no data and functions as a consolidator of locally trained models. The aggregator node is specifically designed to possess a Representational State Transfer (REST)–application programming interface (API) termed as the API Forwarder. The API Forwarder is responsible for managing the requests received from the data stations and subsequently routing them to the corresponding active Docker container, running the aggregation algorithm.

To prevent any malicious or unauthorized communication with the aggregator node, each data station is equipped with a JSON Web Token (JWT) that is unique for each iteration. The API Forwarder only accepts communications that are accompanied by a valid JWT. The implementation of this functionality guarantees the protection of infrastructure users and effectively mitigates the risk of unauthorized access to SAS. Figure 3 shows the architecture and execution mechanism for the SAS.

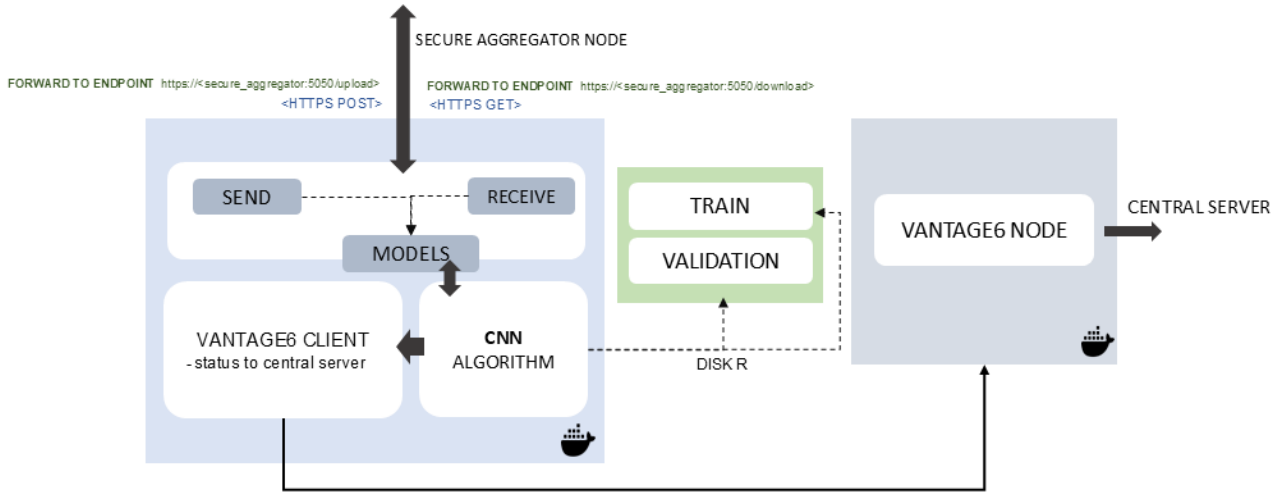
Figure 3. Architecture of the secure aggregation server, showing incoming and outgoing requests from the data station nodes. The upload and download folders are temporary locations used within the running Docker container to store the local and averaged models through disk read or write operations. The API forwarder, running at port 5050 and embedded within the Vantage6 infrastructure, forwards the incoming requests from the data station nodes to the algorithm API running at local port 7000 within the Docker container through HTTP requests. The SAS is hosted behind the firewall of a proxy server, which allows only hypertext transfer protocol secure (HTTPS) communication from the participating nodes. API: application programming interface; FedAvg: federated averaging; JWT: JSON Web Token.



Data Stations

Data stations are devices located within the confines of each hospital's jurisdiction that are not reachable or accessible from external sources other than Vantage6. The data stations communicate with the central server through a pull mechanism. Furthermore, the data stations not only serve as hosts for the infrastructure node but also offer the essential computational resources required for training the deep learning network. The infrastructure node is the software component installed in the data stations that orchestrates the local execution of the model and its communication with the central server and the SAS. Each data station is equipped with at least 1 graphics processing unit (GPU), which enables the execution of CNNs. Preprocessing of the raw CT images was executed locally, using automated preprocessing scripts packaged as Docker containers, and the preprocessed CT images are stored within a file system volume in each station. The CNN Docker is designed and allowed to access the preprocessed images during training. The primary function of the data station is to receive instructions from both the SAS and the central server, perform the computations needed for training the CNN algorithm, and subsequently transmit the model weights back to the respective sources. Figure 4 depicts the architectural layout of the data station and node component of the infrastructure.

Figure 4. Architecture of the data station node component. The node runs the CNN algorithm to learn from the local data. The node further sends and receives model weights from the secure aggregation server. The train and validation folders are persistent locations within the data stations, storing the preprocessed NIFTI images. At the end of each training cycle, the intermediate averaged model is first evaluated on the validation sample. CNN: convolutional neural network; HTTPS: hypertext transfer protocol secure; NIFTI: neuroimaging informatics technology initiative.



Train

The "train" in the form of a Docker image encompasses several components bundled together: an untrained U-Net [48,49], a type of CNN architecture designed for image segmentation tasks for training on local data; the aggregation algorithm used for consolidating the models; and a secondary Python Flask API known as the Algorithm API for facilitating the communication of these models. The Algorithm API is designed to cater to requests from the API Forwarder and is built within the algorithm container. Two levels of API ensured that the node could handle multiple requests and divert to appropriate Docker containers. Furthermore, the first level of API also helps in restricting malicious requests by checking the JWT token signature, so that the models within the master Docker container are protected. Each data station is responsible for training and transmitting the CNN model to the aggregator server. This suggests that the aggregation algorithm exhibits a waiting period during which it ensures that all data stations have effectively transmitted their models to the server before proceeding to the next iterations. The process is executed in an iterative manner until convergence is achieved or the specified number of iterations is attained.

Tracks and Track Provider

The various infrastructure components establish coordination among themselves through the use of secure communication channels commonly referred to as the "tracks." The communication channels are enabled with end-to-end encryption. The responsibility for the maintenance of the infrastructure, including the hosting of the central coordinating server and the specialized SAS, lies with the track provider. The track provider is additionally accountable for the maintenance of the "tracks" and aids the data providers in establishing the local segment of the infrastructure known as the "nodes."

Data Provider

Data providers refer to hospitals and health care organizations that are responsible for curating the pertinent datasets used for training the deep learning network. The responsibility of hosting the data stations within their respective local jurisdiction lies with the data provider. They exercise authority over the data as well as the infrastructure component called the node.

Researcher

The researcher is responsible for activating the deep learning algorithm and engaging in the authentication process with the central coordinating server using a registered username and password. This allows the researcher to establish their identity and gain secure access to the system, with their communication safeguarded through end-to-end encryption. The researcher can then assign tasks to individual nodes, monitor progress, and terminate tasks in the event of failure. Importantly, the researcher's methodology is designed to keep the intermediate outcomes of the iterative deep learning training process inaccessible, ensuring that the ultimate global model can only be obtained upon completion of all training iterations, thereby mitigating the risk of unauthorized access by malicious researchers to the intermediate models and providing a security mechanism against insider attacks.

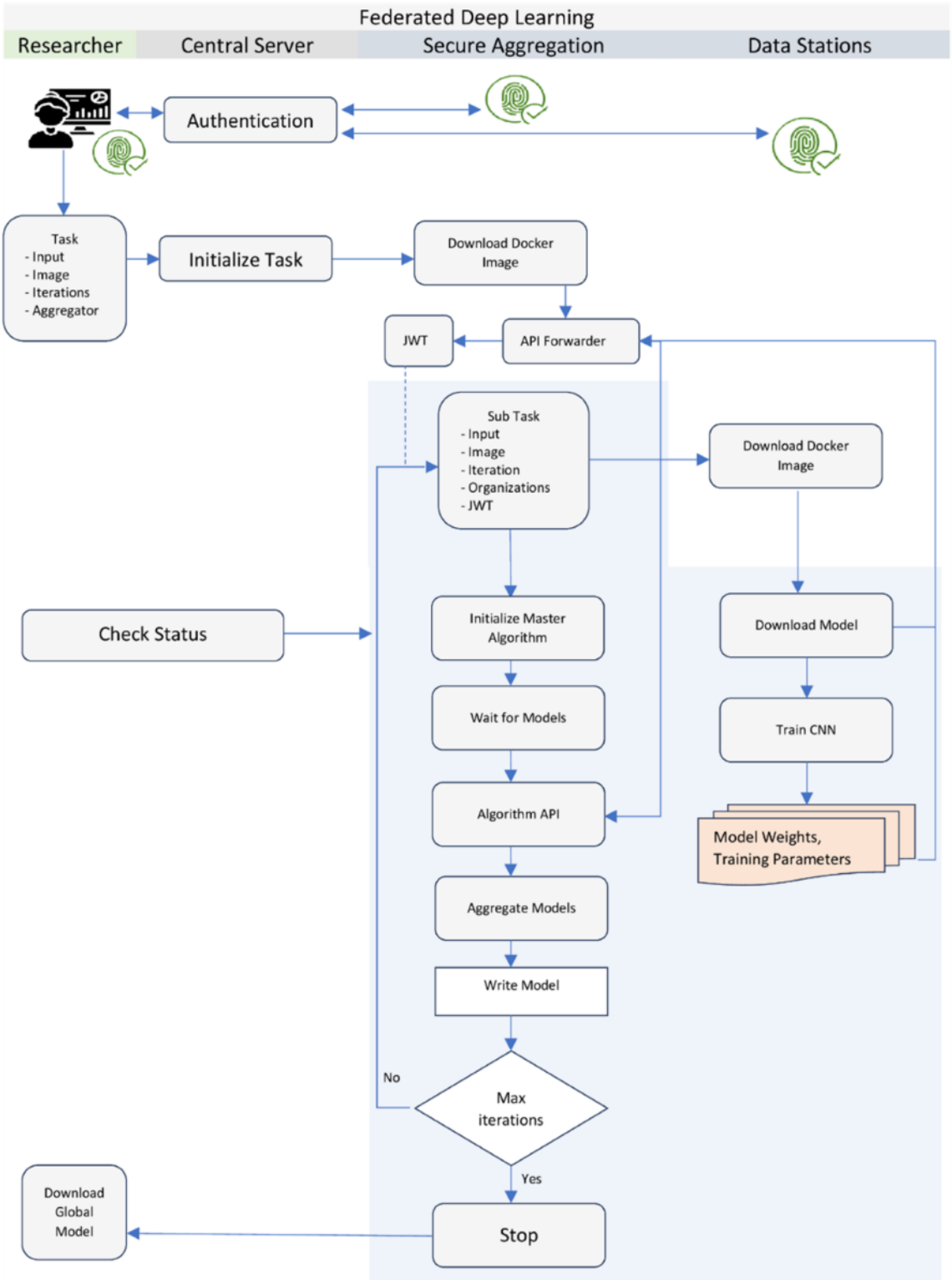
Training Process

Each of the components described above works in a coordinated manner to accomplish the convergence of the deep learning algorithm. The training process begins with the researcher authenticating with the central server. Upon successful authentication, the researcher specifies the task details, including a prebuilt Docker image, input parameters, number of iterations, and the identity of the SAS. The task is then submitted to the central server, which forwards it to the connected nodes. The SAS is the first to receive the task request. It downloads the specified Docker image from the registry and initiates the master algorithm. The master algorithm orchestrates the training at each data station node through the central server. The central server then forwards a subtask request to all the data stations. Like the SAS, the data nodes download the same Docker image and initiate the node part of the algorithm. The node algorithm runs the learning process on local data for the specified number of epochs. After each training cycle, the node algorithm sends the local model weights to the SAS.

The SAS verifies the JWT signature of each received model and forwards the request to the Algorithm API. The Algorithm API extracts the weight and metadata information of the models. Once the SAS receives all the required locally trained models for that cycle, it initiates the FedAvg algorithm to consolidate the models and create an intermediate averaged model, which is stored locally. This completes the first iteration of the training cycle. For the second and subsequent iterations, the data stations request the SAS to send the intermediate averaged model weights from the previous iteration. The SAS validates these requests and sends the model weights to the data stations, which then use them for further training on their local data. This cycle of training and averaging continues until the model converges or the desired number of iterations is reached.

At the end of the training process, the SAS sends a notification to the researcher indicating the successful completion of the task. The researcher can then download the final global model from the server. It is important to note that during the training iterations, the researcher or other users of the infrastructure do not have access to the intermediate averaged models generated by the SAS. This design choice prevents the possibility of insider attacks and data leakage, as users cannot regenerate patterns from the training data using the intermediate models. Figure 5 shows the diagrammatic representation of the training process spread across the infrastructure components.

Figure 5. Process illustration of federated deep learning training. All entities, including the researcher, the central aggregation server, and the data stations, first authenticate with the central server. The researcher creates a task description and submits the task to the central server, which then forwards the request to the secure aggregation node to start the master task. The master task then sends a request to all data stations to download the algorithm Docker image and start training on the local data. Researchers can monitor the algorithm's execution status on the central server using the "check status" function, which reports whether each iteration is completed or aborted as processed by the secure aggregation server and data stations. At the end of each local training, the data stations send the models to the API forwarder of the secure aggregation node by authenticating against a valid JWT token. The JWT token ensures that no unauthorized data station is able to send or receive models from the secure aggregation server. API: application programming interface; CNN: convolutional neural network; JWT: JSON Web Token.



Code Availability

The federated deep learning infrastructure and the algorithm used in this research are open source and publicly available. The codebase, encompassing the components of the infrastructure, the algorithm, and wrappers for running it in the infrastructure and the researcher notebooks, are all available and deposited on GitHub, a public repository platform, under the Apache 2.0 license. This open access allows the research community to scrutinize and leverage our implementation for further development in the field of FL.

The Vantage6 (version 2.0.0) [27,50] open-source software was customized to cater to the specific requirements for running the deep learning algorithm. The central server (Vantage6 version 2.0.0) and the aggregator server were hosted by Medical Data Works BV in 2 separate cloud machines (Microsoft Azure). At each participating center, the "node" component of the software was installed and setup either on a physical or cloud machine running Ubuntu (version 16.0) or above with an installation of Python, (version 3.7 or above; Python Software Foundation), Docker Desktop (personal edition), and NVIDIA CUDA GPU interface (version 11.0). The source code of the customized "node" [51] and setup instructions [52] are available on respective GitHub repositories. The federated deep learning algorithm was adapted to the infrastructure as Python scripts [53] and wrapped in a Docker container. Separately, the "researcher" notebooks [54] containing python scripts for connecting to the infrastructure and running the algorithms are also available on GitHub. Table 2 provides an outline of the resource requirement and computational cost of the experiment.

Table 2. Resource requirement and computational cost.

End points	Resource requirement		Average execution time (per iteration)
	Software	Hardware	
Central server	• Ubuntu (version 16) and above • Docker Desktop • Python (3.7 or above) • Vantage6 (version 2.0.0)	• 4 CPUsa • 16 GB RAM • 20 GB Disk Space	N/A[b]
Data station	• Ubuntu (version 16) and above • Docker Desktop • Python (3.7 or above) • Vantage6 (version 2.0.0) • CUDA GPU Interface (version 11.0)	• 4 CPUs • 1 GPUc • 16 GB RAM • 40 GB disk space	40 mins
Secure aggregation server	• Ubuntu (version 16) and above • Docker Desktop • Python (3.7 or above) • Vantage6 (version 2.0.0)	• 4 CPUs • 16 GB RAM • 40 GB disk space	60 seconds

[a]CPU: central processing unit.

[b]Not applicable.

[c]GPU: graphics processing unit.

Ethical Considerations

The work was performed independently with the ethics board's approval from each participating institution. Approvals from each of the participating institutions including soft copies of approval have been submitted to the leading partner. The lead partner's institutional review board approval (MAASTRO Clinic, The Netherlands) is "W 20 11 00069" (approved on November 24, 2020). The authors attest that the work was conducted by the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975.

Results

Overview

The study was carried out and concluded in 4 primary stages using an agile approach as follows: planning, design and development, partner recruitment, and execution of federated deep learning. The planning phase of the study, which encompassed a meticulous evaluation and determination of the following inquiries, held equal significance to the description of the clinical issue and data requirements.

- What are the minimum resource requirements for each participating center?
- How to design a safe and robust infrastructure to effectively address the requirements of a federated deep learning study?
- How can a reliable and data-agnostic federated deep learning algorithm be designed?
- What are the operational and logistical challenges associated with conducting a large-scale federated deep learning study?

The second phase, that is, the design and development phase, primarily focused on the creation, testing, and customization of the Vantage6 infrastructure for studies specifically focused on deep learning. To meet the security demands of these

investigations, this study involved the development of the SAS, which was not originally included in the Vantage6 architecture. The CNN algorithm was packaged as a Docker container and made compatible with the Vantage6 infrastructure, allowing it to be easily deployed and used within the Vantage6 ecosystem. Prior to the deployment of the algorithm, it underwent testing using multiple test configurations consisting of data stations that were populated with public datasets.

The primary objective of the third phase entailed the recruitment of partners who displayed both interest and suitability from various global locations. The project consortium members became part of the project by obtaining the necessary institutional review board approvals and signing an infrastructure user agreement. This agreement enabled them to install the required infrastructure locally and carry out algorithmic execution. The inclusion criteria for patient data, as well as the technology used for data anonymization and preprocessing, were provided to each center. The team collaborated with each partner center to successfully implement the local component of the infrastructure.

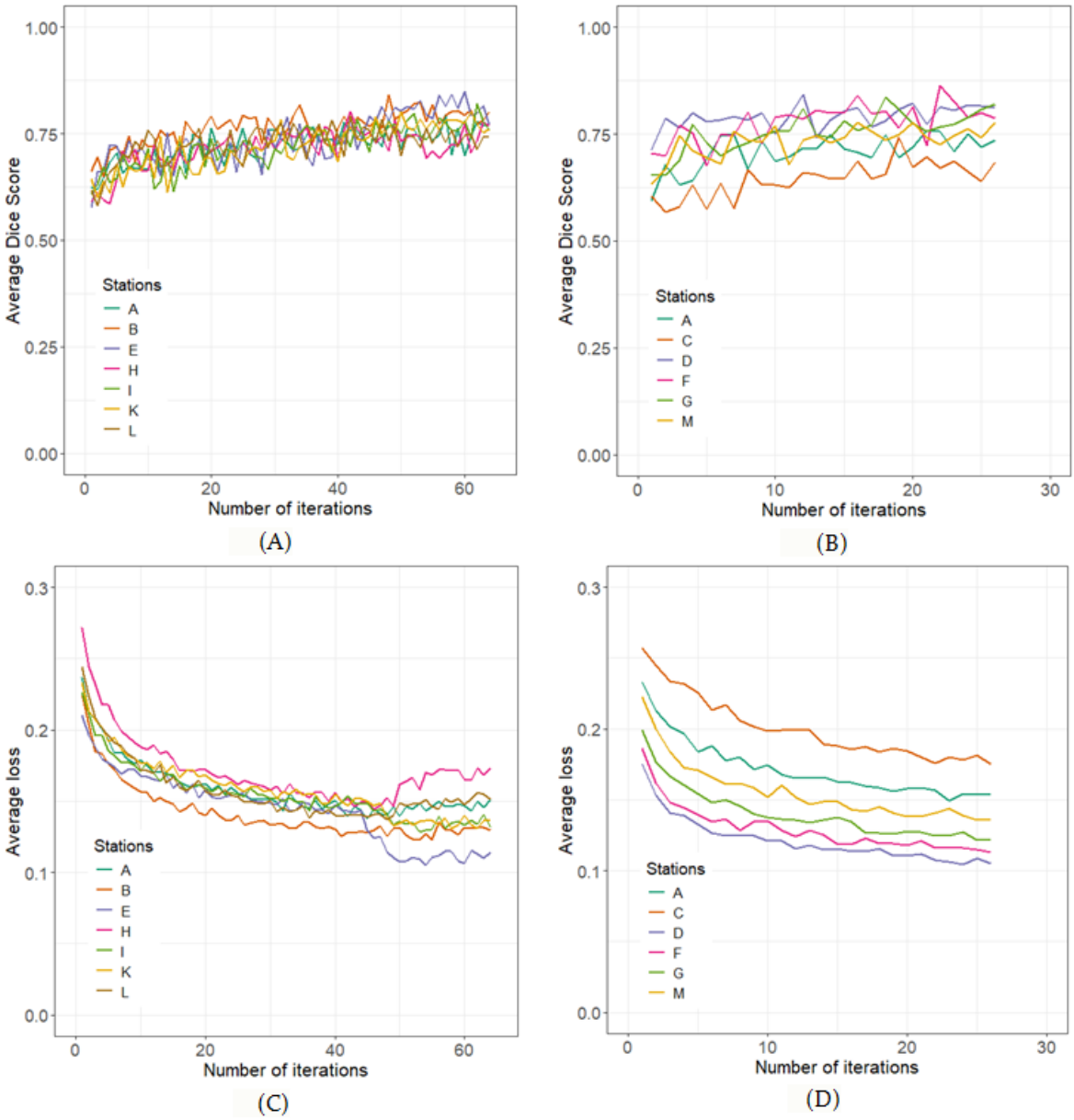
The concluding stage of the study involved the simultaneous establishment of connections between all partner centers and the existing infrastructure. The algorithm was subsequently initiated by the researcher and the completion of the predetermined set of federated iterations was awaited across all centers.

Proof of Concept

The architectural strategy described above was implemented among ARGOS consortium partners on real-world lung cancer CT scans. For an initial “run-up” of the system, we deployed the abovementioned PHT system across 12 institutions, located in 8 countries and 4 continents. A list of members participating in the ARGOS consortium can be found on the study protocol [26]. In total, 2078 patients’ data were accessible via the infrastructure for training (n=1606) and holdout validation (n=472). For this initial training experiment, the 12 centers were divided into 2 groups. The first, referred to as group A, comprised 7 collaborators, and we were able to reach a total of 64 iterations of model training each with 10,000 steps per iteration. Likewise, group B comprising 6 hospitals was able to train the deep learning model for 26 iterations. It was observed that no significant improvement of the model was observed for both groups after 26th iteration. The results from the proof-of-concept study are shown in Figure 6.

While the training time for the models was similar at each center, how quickly they could be uploaded and downloaded depended heavily on the quality of the internet connection. This meant the entire process was significantly slowed down by the center with the slowest internet.

Figure 6. Plots showing the results from training the convolutional neural network on two groups as follows: group 1 (A, B, E, H, I, K, L) and group 2 (A, C, D, F, G, M). (A) Average Dice score per iteration of the model trained on group 1. (B) Average Dice score per iteration of the model trained on group 2. (C) Average training loss per iteration of the model trained on group 1. (D) Average training loss per iteration of the model trained on group 2.



Discussion

This study demonstrated the feasibility of a privacy-preserving federated deep learning infrastructure and presented a proof-of-concept study for GTV segmentation in patients with lung cancer. Using the PHT framework, the infrastructure linked 12 hospitals across 8 nations, showcasing its scalability and global applicability. Notably, throughout the process, no patient data were shared outside the participating institutions, addressing significant data privacy concerns. The introduction of a SAS further ensured that model averaging occurred in a secure environment, mitigating potential data leakage issues in FL.

One of the most used methodologies in recent years has been the use of FL for promoting research on privacy-sensitive data. To orchestrate FL on nonstructured data in the horizontal partitioning context, it is essential to develop specialized software for edge computation and technical infrastructures for cloud aggregation. These infrastructures enable federated machine learning (FML) responsibilities to be carried out in a secure and regulated manner. However, only a limited number of these studies have documented the background governance strategies and the ethical, legal, and social implications framework for conducting such studies.

The study presented a novel approach for executing large-scale federated deep learning on medical imaging data, integrating geographically dispersed real-world patient data from cross-continental hospital sites. The deep learning algorithm was designed to automatically delineate the GTV from chest CT images of patients with lung cancer who underwent radiotherapy treatment. The underlying FL infrastructure architecture was designed to securely perform deep learning training and was tested for vulnerabilities from known security

threats. This paper predominantly discussed the FL infrastructure architecture and presented a firsthand experience of conducting such studies. The preliminary training of the deep learning algorithm serves as the feasibility demonstration of the methodology, and further refinement is required to achieve acceptable clinical-grade accuracy and generalizability.

The study used an open-source and freely accessible technological stack to demonstrate the feasibility and applicability of federated deep learning. Vantage6, a Python-based FL infrastructure, is used to train and coordinate deep learning execution. TensorFlow and Flask, both open-source Python libraries, are used for the development of the algorithm, subsequently encapsulated within Docker services for containerization purposes. The communication channels between the hospital, central server, and the aggregation node have been secured using Hypertext Transfer Protocol Secure and Secure Hash Algorithm encryption. The hospital sites' computer systems were based on the Ubuntu operating system and equipped with at least 1 GPU to enhance computational capabilities. The participating centers had the flexibility to choose any CUDA-compatible GPU devices and determine the number of GPUs to use, enabling resource-constrained centers to contribute. However, a limitation exists in terms of computational time due to the synchronous training process being dependent on the slowest participant.

The infrastructure has been tested against known security attacks and as defined by the Open Worldwide Application Security Project top-ten categories [55]. It has been found that the Vantage6 app is impeccable against insecure design, software and data integrity failures, security logging and monitoring failures, and server-side request forgery and sufficiently secured against broken access control, cryptographic failures, injection, security misconfigurations, vulnerable and outdated components, and finally identification and authentication failures. Since the infrastructure is dependent on other underlying technologies like Docker and Flask-API, the security measures in these technologies also affect the overall security of the infrastructure. Additionally, the infrastructure is hosted behind proxy firewalls, adding to its overall security against external threats.

In this study, we implemented a SAS positioned between the data nodes (eg, hospitals and clinics) and the central server. The SAS plays a crucial role in strengthening the privacy and confidentiality of the learning process. The SAS acts as an intermediary that temporarily stores the local model updates from the participating data nodes, ensuring complete isolation from the central server, researchers, and any external intruders. The key benefits of using a dedicated SAS over a random aggregation mechanism in FL are as follows:

- Privacy protection of individual user data and model updates:
  - The secure aggregation protocol ensures that the central server only learns the aggregated sum of all user updates, without being able to access or infer the individual user's private data or model updates.
  - By isolating the intermediate updates, the secure aggregation process prevents external attackers from performing model inversion attacks.
- Tolerance to user dropouts:
  - The SAS is designed to handle situations where some users fail to complete the execution. In the case of synchronous training, the server stores the latest successful model, enabling data nodes to pick up where they left off instead of restarting from scratch.
- Integrity of the aggregation process:
  - The secure aggregation protocol provides mechanisms to verify the integrity of the intermediate models by allowing only the known data nodes to send a model. This maintains the reliability and trustworthiness of the FL system.

FL offers 2 main approaches for model aggregation: sending gradients or weights [56,57]. In gradient sharing, data nodes update local models and transmit the gradients of their parameters for aggregation. Conversely, weight sharing involves sending the fully updated model weights directly to the server for aggregation. Sharing gradients have a higher risk of model inversion attacks. In the study presented here, the data nodes sent model weights instead of model gradients, thus preventing the "gradient leakage" problem. However, weight sharing is not failproof either [58], and the SAS plays a crucial role again in preventing users—internal or external—from accessing the weights from the aggregator machine.

The deployment of the FL infrastructure and training of the deep learning algorithm presented unique challenges that needed to be catered to. Some of them are listed below:

- Heterogeneity across hospitals: Initially, it was not possible to confirm the technology environment at each site. This required significant work to overcome the obstacles connected with each center while deploying a functional infrastructure, good communication, and efficient algorithms.
- Inconsistent IT policies: Standardizing the setup across institutions was hindered by varying IT governance and network regulations in different health care systems across different countries.
- Clinical expertise gap: The predominance of medical personnel over IT specialists at participating hospitals necessitated extensive documentation to ensure clinician comprehension of the FL process.
- Network bottlenecks: Network configurations at participating sites significantly impacted training duration, often leading to delays in model convergence.

The study presented in the paper has identified several areas that require further investigation and improvement. While the findings are valuable, the infrastructure, algorithm, and processes still need to be made more secure, private, trustworthy, robust, and seamless [59]. For example, incorporating homomorphic encryption of the learned models will enhance privacy and provide model obfuscation against inversion attacks. Finally, to further enhance confidence and trust in federated artificial intelligence, it is crucial to conduct additional studies involving a larger number of participating centers and a thorough clinical evaluation of the models.

Acknowledgments

We would like to express our sincere appreciation and gratitude to Integraal Kankercentrum Nederland (IKNL), the Netherlands, for their invaluable contribution in providing us with the necessary infrastructure support. We express our gratitude to Medical Data Works, the Netherlands, for their role as the infrastructure service provider in hosting the central and secure aggregation server. We also express our gratitude to Varsha Gouthamchand and Sander Puts for their contribution to the successful execution of the experiments. In conclusion, we express our gratitude to the various data-providing organizations for their substantial support and collaboration throughout all stages of the project. AC, LV, RF, and LW acknowledge financial support from the Dutch Research Council (NWO) (TRAIN project, dossier 629.002.212) and the Hanarth Foundation.

Conflicts of Interest

Dr AD and JvS are both cofounders, shareholders, and directors of Medical Data Works B.V.

References

1. Sun C, Ippel L, Dekker A, Dumontier M, van Soest J. A systematic review on privacy-preserving distributed data mining. Data Sci 2021 Oct;4(2):121-150. [doi: 10.3233/DS-210036]
2. Choudhury A, Sun, C, Dekker M, Dumontie J, van Soest. Privacy-preserving federated data analysis: data sharing, protection, bioethics in healthcare. In: El Naqa I, Murphy MJ, editors. Machine Deep Learning in Oncology. Cham, Switzerland: Springer International Publishing; 2022:135-172.
3. Deist TM, Dankers FJ, Ojha P, Scott Marshall M, Janssen T, Faivre-Finn C, et al. Distributed learning on 20 000+ lung cancer patients - the personal health train. Radiother Oncol 2020;144:189-200 [FREE Full text] [doi: 10.1016/j.radonc.2019.11.019] [Medline: 31911366]
4. Choudhury A, Theophanous S, Lønne PI, Samuel R, Guren MG, Berbee M, et al. Predicting outcomes in anal cancer patients using multi-centre data and distributed learning - a proof-of-concept study. Radiother Oncol 2021;159:183-189 [FREE Full text] [doi: 10.1016/j.radonc.2021.03.013] [Medline: 33753156]
5. Beyan O, Choudhury A, van Soest J, Kohlbacher O, Zimmermann L, Stenzhorn H, et al. Distributed analytics on sensitive medical data: the personal healt train. Data Intell 2020;2(1-2):96-107 [FREE Full text] [doi: 10.1162/dint_a_00032]
6. Moncada-Torres A, Martin F, Sieswerda M, van Soest J, Geleijnse G. VANTAGE6: an open source privacy preserving federated learning infrastructure for secure insight exchange. AMIA Annu Symp Proc 2020;2020:870-877 [FREE Full text] [Medline: 33936462]
7. Becker R, Chokoshvili D, Comandé G, Dove ES, Hall A, Mitchell C, et al. Secondary use of personal health data: when is it “Further Processing” under the GDPR, and what are the implications for data controllers? Eur J Health Law 2022;30(2):129-157. [doi: 10.1163/15718093-bja10094]
8. El Naqa I, Ruan D, Valdes G, Dekker A, McNutt T, Ge Y, et al. Machine learning and modeling: data, validation, communication challenges. Med Phys 2018;45(10):e834-e840 [FREE Full text] [doi: 10.1002/mp.12811] [Medline: 30144098]
9. van Stiphout R. How to share data and promote a rapid learning health medicine? In: Valentini HJ, Schmoll C, van de Velde JH, editors. Multidisciplinary Management of Rectal Cancer. Cham, Switzerland: Springer International Publishing; 2018:623-634.
10. Kazmierska J, Hope A, Spezi E, Beddar S, Nailon WH, Osong B, et al. From multisource data to clinical decision aids in radiation oncology: the need for a clinical data science community. Radiother Oncol 2020;153:43-54 [FREE Full text] [doi: 10.1016/j.radonc.2020.09.054] [Medline: 33065188]
11. Fischer-Hübner S. Privacy-enhancing technologies. In: Liu T, Özsu MT, editors. Encyclopedia of Database Systems. Boston, MA: Springer; 2009:2142-2147.
12. Coopamootoo KPL. Usage patterns of privacy-enhancing technologies. In: ACM Digital Library. 2020 Presented at: CCS '20: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security; November 2, 2020; New York, NY URL: https://dl.acm.org/doi/10.1145/3372297.3423347
13. Emerging privacy-enhancing technologies. OECD. URL: https://www.oecd.org/publications/emerging-privacy-enhancing-technologies-bf121be4-en.htm [accessed 2025-04-25]
14. Kavianpour S, Sutherland J, Mansouri-Benssassi E, Coull N, Jefferson E. Next-generation capabilities in trusted research environments: interview study. J Med Internet Res 2022;24(9):e33720 [FREE Full text] [doi: 10.2196/33720] [Medline: 36125859]
15. Design a secure research environment for regulated data. Microsoft. URL: https://learn.microsoft.com/en-us/azure/architecture/ai-ml/architecture/secure-compute-for-research [accessed 2024-04-25]
16. Imaging data commons. National Cancer Institute Cancer Research Data Commons. URL: https://datacommons.cancer.gov/repository/imaging-data-commons [accessed 2024-04-25]

17. Kotter E, Marti-Bonmati L, Brady AP, Desouza NM. ESR white paper: blockchain and medical imaging. Insights Imaging 2021;12(1):82 [FREE Full text] [doi: 10.1186/s13244-021-01029-y] [Medline: 34156562]
18. Sultana M, Hossain A, Laila F, Taher KA, Islam MN. Towards developing a secure medical image sharing system based on zero trust principles and blockchain technology. BMC Med Inform Decis Mak 2020;20(1):256 [FREE Full text] [doi: 10.1186/s12911-020-01275-y] [Medline: 33028318]
19. Manifesto of the personal health train consortium. Data Driven Life Sciences. URL: https://www.dtls.nl/wp-content/uploads/2017/12/PHT_Manifesto.pdf [accessed 2024-03-11]
20. McMahan E, Moore D, Ramage S, Hampson BA. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of Machine Learning Research. 2017 Presented at: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics; April 20-22, 2017; Fort Lauderdale, FL URL: https://proceedings.mlr.press/v54/mcmahan17a.html
21. Zhang C, Choudhury A, Shi Z, Zhu C, Bermejo I, Dekker A, et al. Feasibility of privacy-preserving federated deep learning on medical images. Int J Radiat Oncol Biol Phys 2020;108(3):e778. [doi: 10.1016/j.ijrobp.2020.07.234]
22. Choudhury A, van Soest J, Nayak S, Dekker A. Personal health train on FHIR: a privacy preserving federated approach for analyzing FAIR data in healthcare. In: Bhattacharjee A, Kr. Borgohain S, Soni B, Verma G, Gao XZ, editors. Machine Learning, Image Processing, Network Security and Data Sciences. Singapore: Springer; 2020.
23. Gouthamchand V, Choudhury A, P Hoebers FJ, R Wesseling FW, Welch M, Kim S, et al. Making head and neck cancer clinical data findable-accessible-interoperable-reusable to support multi-institutional collaboration and federated learning,? BJR Artif Intell 2024;1(1).
24. Sun C, van Soest J, Koster A, Eussen SJ, Schram MT, Stehouwer CD, et al. Studying the association of diabetes and healthcare cost on distributed data from the maastricht study and statistics Netherlands using a privacy-preserving federated learning infrastructure. J Biomed Inform 2022;134:104194 [FREE Full text] [doi: 10.1016/j.jbi.2022.104194] [Medline: 36064113]
25. Railway governance. Medical Data Works. URL: https://www.medicaldataworks.nl/governance [accessed 2024-09-11]
26. Dekker A. ARtificial Intelligence for Gross Tumour vOlume Segmentation (ARGOS). National Library of Medicine. URL: https://clinicaltrials.gov/study/NCT05775068 [accessed 2024-01-11]
27. Overview: what is vantage6? Vantage6 documentation. URL: https://docs.vantage6.ai/en/main/ [accessed 2024-04-11]
28. U-Net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham, Switzerland: Springer; Nov 18, 2015.
29. Patil RB. Prognostic and prediction modelling with radiomics for non-small cell lung cancer. Maastricht University. 2020. URL: https://cris.maastrichtuniversity.nl/en/publications/prognostic-and-prediction-modelling-with-radiomics-for-non-small- [accessed 2020-10-06]
30. Tao Z, Lyu S. A survey on automatic delineation of radiotherapy target volume based on machine learning. Data Intell 2023;5(3):814-856. [doi: 10.1162/dint_a_00204]
31. Liu X, Li KW, Yang R, Geng LS. Review of deep learning based automatic segmentation for lung cancer radiotherapy. Front Oncol 2021;11:717039 [FREE Full text] [doi: 10.3389/fonc.2021.717039] [Medline: 34336704]
32. Ma Y, Mao J, Liu X, Dai Z, Zhang H, Zhang X, et al. Deep learning-based internal gross target volume definition in 4D CT images of lung cancer patients. Med Phys 2023;50(4):2303-2316. [doi: 10.1002/mp.16106] [Medline: 36398404]
33. Zhang F, Wang Q, Li H. Automatic segmentation of the gross target volume in non-small cell lung cancer using a modified version of ResNet. Technol Cancer Res Treat 2020;19:153303382094748. [doi: 10.1177/1533033820947484]
34. Xie H, Chen Z, Deng J, Zhang J, Duan H, Li Q. Automatic segmentation of the gross target volume in radiotherapy for lung cancer using transresSEUnet 2.5D network. J Transl Med 2022;20(1):524 [FREE Full text] [doi: 10.1186/s12967-022-03732-w] [Medline: 36371220]
35. Raimondi D, Chizari H, Verplaetse N, Löscher BS, Franke A, Moreau Y. Genome interpretation in a federated learning context allows the multi-center exome-based risk prediction of Crohn's disease patients. Sci Rep 2023 Nov 09;13(1):19449 [FREE Full text] [doi: 10.1038/s41598-023-46887-2] [Medline: 37945674]
36. Riedel P, von Schwerin R, Schaudt D, Hafner A, Späte C. ResNetFed: federated deep learning architecture for privacy-preserving pneumonia detection from COVID-19 chest radiographs. J Healthc Inform Res 2023;7(2):203-224 [FREE Full text] [doi: 10.1007/s41666-023-00132-7] [Medline: 37359194]
37. Nazir S, Kaleem M. Federated learning for medical image analysis with deep neural networks. Diagnostics (Basel) 2023;13(9):1532 [FREE Full text] [doi: 10.3390/diagnostics13091532] [Medline: 37174925]
38. Shiri I, Vafaei Sadr A, Akhavan A, Salimi Y, Sanaat A, Amini M, et al. Decentralized collaborative multi-institutional PET attenuation and scatter correction using federated deep learning. Eur J Nucl Med Mol Imaging 2023;50(4):1034-1050 [FREE Full text] [doi: 10.1007/s00259-022-06053-8] [Medline: 36508026]
39. Zhang M, Qu L, Singh P, Kalpathy-Cramer J, Rubin DL. SplitAVG: a heterogeneity-aware federated deep learning method for medical imaging. IEEE J Biomed Health Inform 2022;26(9):4635-4644. [doi: 10.1109/jbhi.2022.3185956]
40. Shiri I, Vafaei Sadr A, Amini M, Salimi Y, Sanaat A, Akhavanallaf A, et al. Decentralized distributed multi-institutional PET image segmentation using a federated deep learning framework. Clin Nucl Med 2022;47(7):606-617. [doi: 10.1097/rlu.0000000000004194]

41. Sarma KV, Harmon S, Sanford T, Roth HR, Xu Z, Tetreault J, et al. Federated learning improves site performance in multicenter deep learning without data sharing. J Am Med Inform Assoc 2021;28(6):1259-1264 [FREE Full text] [doi: 10.1093/jamia/ocaa341] [Medline: 33537772]
42. Harmon SA, Sanford TH, Xu S, Turkbey EB, Roth H, Xu Z, et al. Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nat Commun 2020;11(1):4080 [FREE Full text] [doi: 10.1038/s41467-020-17971-2] [Medline: 32796848]
43. Durga R, Poovammal E. FLED-block: federated learning ensembled deep learning blockchain model for COVID-19 prediction. Front Public Health 2022;10:892499 [FREE Full text] [doi: 10.3389/fpubh.2022.892499]
44. Pati S, Baid U, Edwards B, Sheller M, Wang S, Reina GA, et al. Federated learning enables big data for rare cancer boundary detection. Nat Commun 2022;13(1):7346 [FREE Full text] [doi: 10.1038/s41467-022-33407-5] [Medline: 36470898]
45. Leroy V, Ananya C, Aiara LG, Andre D, Leonard W. Feasibility of training federated deep learning oropharyngeal primary tumor segmentation models without sharing gradient information. Research Square Preprint published online 25 July, 2024 [FREE Full text] [doi: 10.21203/rs.3.rs-4644605/v1]
46. Schmidt K, Bearce B, Chang K, Coombs L, Farahani K, Elbatel M, et al. Fair evaluation of federated learning algorithms for automated breast density classification: the results of the 2022 ACR-NCI-NVIDIA federated learning challenge. Med Image Anal 2024;95:103206. [doi: 10.1016/j.media.2024.103206] [Medline: 38776844]
47. Pati S, Kumar S, Varma A, Edwards B, Lu C, Qu L, et al. Privacy preservation for federated learning in health care. Patterns (N Y) 2024;5(7):100974 [FREE Full text] [doi: 10.1016/j.patter.2024.100974] [Medline: 39081567]
48. Oreiller V, Andrearczyk V, Jreige M, Boughdad S, Elhalawani H, Castelli J, et al. Head and neck tumor segmentation in PET/CT: the HECKTOR challenge. Med Image Anal 2022;77:102336 [FREE Full text] [doi: 10.1016/j.media.2021.102336] [Medline: 35016077]
49. Iantsen A, Jaouen V, Visvikis D, Hatt M. Squeeze-and-excitation normalization for brain tumor segmentation. In: Crimi A, Bakas S, editors. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Cham, Switzerland: Springer International Publishing; 2021.
50. IKNL/vantage6: Docker CLI package for the vantage6 infrastructure. GitHub. URL: https://github.com/IKNL/vantage6/tree/DEV3 [accessed 2024-05-01]
51. Martin F. Featured communities. Zenodo. URL: https://doi.org/10.5281/zenodo.3686944 [accessed 2024-05-06]
52. MaastrichtU-CDS/argos-infrastructure. GitHub. URL: https://github.com/MaastrichtU-CDS/argos-infrastructure [accessed 2024-05-01]
53. MaastrichtU-CDS/projects_argos_argos-code-repo_full-algorithm. GitHub. URL: https://github.com/MaastrichtU-CDS/projects_argos_argos-code-repo_full-algorithm [accessed 2024-05-01]
54. MaastrichtU-CDS/projects_argos_argos-code-repo_researcher-notebooks. GitHub. URL: https://github.com/MaastrichtU-CDS/projects_argos_argos-code-repo_researcher-notebooks [accessed 2024-05-01]
55. OWASP top ten. OWASP Foundation. URL: https://owasp.org/www-project-top-ten/ [accessed 2024-05-02]
56. Moshawrab M, Adda M, Bouzouane A, Ibrahim H, Raad A. Reviewing federated learning aggregation algorithms; strategies, contributions, limitations and future perspectives. Electronics 2023;12(10):2287. [doi: 10.3390/electronics12102287]
57. Liu P, Xu X, Wang W. Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives. Cybersecurity 2022;5(1). [doi: 10.1186/s42400-021-00105-6]
58. Boenisch F, Dziedzic A, Schuster R, Shamsabadi S, Shumailov I, Papernot N. When the curious abandon honesty: federated learning is not private. 2023 Presented at: IEEE 8th European Symposium on Security and Privacy (EuroS&P); July 07, 2023; Delft, theNetherlands. [doi: 10.1109/eurosp57164.2023.00020]
59. Sebastian G, George A, Jackson G. Persuading patients using rhetoric to improve artificial intelligence adoption: experimental study. J Med Internet Res 2023;25:e41430 [FREE Full text] [doi: 10.2196/41430] [Medline: 36912869]

Abbreviations

API: application programming interface
ARGOS: artificial intelligence for gross tumor volume segmentation
CNN: convolutional neural network
CT: computed tomography
FedAvg: federated averaging
FL: federated learning
FML: federated machine learning
GPU: graphics processing unit
GTV: gross tumor volume
HIPAA: Health Insurance Portability and Accountability Act
JWT: JSON Web Token
PHT: Personal Health Train
REST: Representational State Transfer

SAS: secure aggregation server
SRE: secure research environment

Edited by Y Huo; submitted 23.05.24; peer-reviewed by AT Tran, G Sebastian; comments to author 02.07.24; revised version received 01.10.24; accepted 17.10.24; published 06.02.25.

Please cite as:
Choudhury A, Volmer L, Martin F, Fijten R, Wee L, Dekker A, Soest JV
Advancing Privacy-Preserving Health Care Analytics and Implementation of the Personal Health Train: Federated Deep Learning Study
JMIR AI 2025;4:e60847
URL: https://ai.jmir.org/2025/1/e60847
doi:10.2196/60847
PMID:

Original Paper

Predicting Satisfaction With Chat-Counseling at a 24/7 Chat Hotline for the Youth: Natural Language Processing Study

Silvan Hornstein[1], MSc; Ulrike Lueken[1,2], Prof Dr; Richard Wundrack[3], PhD; Kevin Hilbert[4], Prof Dr

[1]Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany

[2]German Center for Mental Health (DZPG), Partner site Berlin/Potsdam, Potsdam, Germany

[3]Krisenchat gGmbH, Berlin, Germany

[4]Department of Psychology, HMU Erfurt - Health and Medical University Erfurt, Erfurt, Germany

Corresponding Author:
Silvan Hornstein, MSc
Department of Psychology
Humboldt-Universität zu Berlin
Wolfgang Köhler-Haus
Rudower Ch 18
Berlin, 12489
Germany
Phone: 49 15753685796
Email: silvan.hornstein@hu-berlin.de

Abstract

Background: Chat-based counseling services are popular for the low-threshold provision of mental health support to youth. In addition, they are particularly suitable for the utilization of natural language processing (NLP) for improved provision of care.

Objective: Consequently, this paper evaluates the feasibility of such a use case, namely, the NLP-based automated evaluation of satisfaction with the chat interaction. This preregistered approach could be used for evaluation and quality control procedures, as it is particularly relevant for those services.

Methods: The consultations of 2609 young chatters (around 140,000 messages) and corresponding feedback were used to train and evaluate classifiers to predict whether a chat was perceived as helpful or not. On the one hand, we trained a word vectorizer in combination with an extreme gradient boosting (XGBoost) classifier, applying cross-validation and extensive hyperparameter tuning. On the other hand, we trained several transformer-based models, comparing model types, preprocessing, and over- and undersampling techniques. For both model types, we selected the best-performing approach on the training set for a final performance evaluation on the 522 users in the final test set.

Results: The fine-tuned XGBoost classifier achieved an area under the receiver operating characteristic score of 0.69 (P<.001), as well as a Matthews correlation coefficient of 0.25 on the previously unseen test set. The selected Longformer-based model did not outperform this baseline, scoring 0.68 (P=.69). A Shapley additive explanations explainability approach suggested that help seekers rating a consultation as helpful commonly expressed their satisfaction already within the conversation. In contrast, the rejection of offered exercises predicted perceived unhelpfulness.

Conclusions: Chat conversations include relevant information regarding the perceived quality of an interaction that can be used by NLP-based prediction approaches. However, to determine if the moderate predictive performance translates into meaningful service improvements requires randomized trials. Further, our results highlight the relevance of contrasting pretrained models with simpler baselines to avoid the implementation of unnecessarily complex models.

Trial Registration: Open Science Framework SR4Q9; https://osf.io/sr4q9

(JMIR AI 2025;4:e63701) doi:10.2196/63701

KEYWORDS

digital mental health; mental illness; mental disorder; adolescence; chat counseling; machine learning; artificial intelligence; large language model; natural language processing; deep learning

Introduction

Most mental health disorders develop early in life [1,2], causing a massive burden on an individual [3], as well as societal, level [4]. This makes early intervention in youth highly relevant [5]. In sharp contrast to the need, accessing help has been described as challenging for young people [5-7]. Therefore, low-threshold services are needed to tackle the burden of mental illness [8].

One such form of intervention gaining popularity is chat-based counseling hotlines [9-11]. Smartphones and chat interactions play a crucial role in youth life [12,13]. The ability to access help within their native digital life reduces numerous health care barriers, making the services a common first access point of help for youth [14]. Indeed, heavy utilization and adoption of those services have been reported globally [14-16]. In addition, the first evidence supports the acceptability [14] and effectiveness [17] of 24/7 chat services.

Considering the increasingly established relevance of those hotlines, the implementation of technological innovation could be highly impactful for the timely and efficient provision of care to youth. Repeatedly, artificial intelligence (AI) has been framed as a key potential for improvements in mental health care [18,19], as well as within digital settings [20]. As AI depends on the availability of large and high-dimensional datasets, chat services seem a quite promising candidate for that. This has indeed been used for diverse natural language processing (NLP) approaches, the subbranch of AI dealing with language. For example, an NLP-based triaging system has been reported to be able to reduce waiting times for those in crisis at a chat hotline [21]. Data-driven decisions regarding further treatment paths have also been investigated by looking into the prediction of recurrent chatting [22] or premature departure from conversations [23]. As suicide risk is a common case at chat hotline services [24], other work focused on early detection and intervention in those situations. Here, several model structures and algorithmic approaches have been suggested [25,26].

This study intends to contribute to the development of NLP approaches within youth chat counseling hotlines. Specifically, the promising but underinvestigated use case of automated evaluation of service quality will be explored. A recent study linked asynchronous chat counseling interactions with reported outcomes and satisfaction of the chatters, using a large dataset of more than 150,000 clients and reporting promising effect sizes of multiple *R*'s of around 0.45 [27]. Another past approach investigated the prediction of chat quality on a label of 675 transcripts of chat counseling sessions [28]. However, while we were not able to find a similar-minded approach within 24/7 hotline services, automated quality evaluation seems particularly relevant for those. Early experiences with help seeking have been linked with future help-seeking behavior in the past [29]. As often being the first contact with any kind of institutionalized help for youth [14], the satisfaction with this interaction is therefore arguably highly relevant for further help-seeking behavior. The reliable identification of those with negative experiences would allow a timely intervention by following up or referrals to other services. Second, the low threshold nature of counseling hotlines makes evaluation more difficult, as it is hard to collect follow-up responses from young help seekers. For example, the aforementioned study of chat hotline effectiveness reported a response rate of 22% among the users [17]. There is also the risk of a bias toward those more satisfied being more likely to respond, which is seen as a common methodical problem in evaluation sciences [30,31]. The ability to estimate the satisfaction with the service out of the conversation data for those who did not respond to any follow-up surveys could therefore significantly improve the evaluation and monitoring of the service quality.

In light of the relevance of the automated evaluation of chat interactions at chat hotlines, as well as the interventions raising relevance for youth mental health care, this project uses a naturalistic sample of 2609 young chatters that were counseled by the German 24/7 hotline service krisenchat. Feedback regarding the perceived helpfulness of the chat is used to train classifiers on the anonymized consultation texts. Performance is evaluated on a previously unseen test set addressing the feasibility of the approach, hypothesizing that we can significantly predict the feedback response by the chatter. Additionally, we assume that applying a pretrained transformer-based model as the state-of-the-art NLP will allow us to outperform a simpler non–transformer-based approach.

Methods

Preregistration

This study was preregistered at Open Science Framework [32]. The preregistration was updated once, as we adapted the used statistical test for the algorithm comparison (see the *Final Evaluation* section under *Methods*) and corrected the questionnaire item used for the outcome variable. We used the checklist for reporting machine learning studies by Klement and El Emam [33], which can be found in Multimedia Appendix 1. Due to legal restrictions regarding the highly vulnerable sample of this study, we are unable to share the dataset. However, the code used for training the algorithm and predicting the helpfulness can be found on GitHub [34], as a starting point for future work.

Ethical Considerations

The data collected and used for this study were part of a larger research project that was ethically approved by the University of Leipzig (372/21-ek). Additionally, we submitted the proposed secondary data analysis to the ethics committee of the Humboldt-Universität zu Berlin. They confirmed that this analysis does not require additional approval. Before the use of this study, the data were subject to a multistep anonymization procedure. Specifically, personally identifying information was marked by counselors and deleted by the organization. Additionally, there also was an automatized method in place to delete names and locations that might have been missed by the counselors. Finally, a k-anonymity principle was applied, deleting all words that were not part of at least 5 different chats.

Setting and Intervention

The anonymized data used for this study were provided by krisenchat, a German 24/7 chat counseling service for people

aged up to 25 years. At krisenchat, those contacting the service through WhatsApp are provided with chat counseling, either by volunteer or employed psychologists, psychotherapists, or social workers. A central aspect of the consultations is the provision of exercises and resources, for example, by sharing YouTube videos, blog posts, or providing them within the chat. However, counselors are also trained in providing emotional support as needed, as well as providing information about mental health care structures in Germany, such as access to psychotherapy or the youth office.

Sample

Data were accessed and shared by the organization on January 17, 2024. On this date, there were feedback questionnaires available for 4560 chatters. Those questionnaires were sent out as part of a larger research project on the service [14]. A total of 264 participants were either younger than 13 years or older than 25 years of age and therefore excluded. While the upper age limit resulted from the scope of the service, the lower age limit resulted from data privacy considerations. An additional 1631 of the chatters were in contact with the service in the last 4 months. A help seeker's inactivity for at least 4 months is an organizational requirement for assuming the consultation purpose has ended and the chat is deleted by anonymization. Accordingly, active chats were also excluded, leading to 2664 concluded conversations and the related feedback questionnaire, with feedback provided between July 22, 2022, and September 17, 2023. For those cases, all messages exchanged between help seekers and counselors within 72 hours before the response to the feedback questionnaire were included. We then excluded cases where conversations consisted of fewer than 10 messages. This led to additional exclusions and resulted in a final sample of 2609 chatters. Their consultations consisted of 141,404 messages, 82,335 by the help seekers and 59,052 by the counselors. Therefore, on average, there were 54 messages exchanged in the three days before the feedback response, 23 messages by the counselor and 31 messages by the help seeker.

Outcome Variable

The feedback questionnaire answered by the chatters included several questions regarding the chat interaction (see Multimedia Appendix 2 for the full questionnaire). For this study, we decided on the use of a single item asking for the helpfulness of the chat ("Did the chat help you?" in German: "Hat dir der Chat geholfen?"), as being the most direct assessment available of chat quality and success, as perceived by the young clients. While the item had four possible answers ("Yes," "Rather Yes," "Rather No," and "No"), we decided to dichotomize it into "Yes" or "No." Reasons for that were improved actionability (as most clinical decision-making is binary by nature, such as providing additional help—yes or no), as well as considering the high-class imbalance. Overall, 89% (n=2332) of the chatters rated the chat as helpful. Specifically, 61 chatters responded with "No," 216 chatters responded with "Rather No," 1138 chatters responded with "Rather Yes," and 1194 chatters responded with "Yes."

Algorithm Training

All decisions regarding algorithmic specifications were made on the 80% of the available data used as a training set. Specifically, we separated the newest 20% of the consultations (522 chats who submitted their feedback after May 27, 2023) as a test set, a commonly used approach to mimic the evaluation of a previously implemented model (eg, [35]).

For our non–transformer-based approach, we preprocessed the data by lowering all words, deleting stop words, and using a lemmanizer [36]. Afterward, a term frequency-inverse document frequency (TF-IDF) vectorizer was used for feature extraction. This vectorizer counts the occurrences of words and weights them based on their frequency across the whole sample. This algorithm was trained using a 5-times repeated 5-fold stratified cross-validation principle. Hyperparameters were tuned using Bayesian optimization maximizing the receiver operating characteristic (ROC) area under the curve (AUC) score for 250 iterations. While there has been some discussion about the applicability of this metric facing class imbalance (eg, [37]), we saw its appropriateness backed up by systematic comparisons [38] and analysis [39] on the issue. All hyperparameters optimized during this procedure are summarized in Table 1. Those also included, as suggested by a reviewer, the range of ngrams used by the vectorizer. Therefore, bigrams and trigrams of words of the messages were also usable as predictors. The used over- or undersampling method was also selected during this procedure, comparing oversampling, undersampling, and Synthetic Minority Oversampling Technique [40]. As a classifier, we applied and tuned an extreme gradient boosting (XGBoost) [41] classifier, as well as a logistic regression. The training pipeline can be found on GitHub.

Table 1. Overview of shortlisted transformer-based models.

Model	Input length, n	Source
uklfr/gottbert-base	512	[42]
distilbert/distilbert-base-german-cased	512	[43]
LennartKeller/longformer-gottbert-base-8192-aw512	8192	[44]

We used hugging face for all transformer-based approaches [42]. We shortlisted GottBERT [43], as well as a German DistilBERT model [44], as language-specific models to be evaluated. However, we assumed that a significant share of our data would exceed those models' input length. Therefore, we also intended to evaluate a Longformer model [45]. This model can process much longer input sequences at reasonable computational costs by applying a sparse attention mechanism (see Table 1 for the shortlisted models including links). We also intended to explore over- and undersampling, as well as class weights to tackle the class imbalance. To represent the chat structure appropriately to the algorithm, we introduced two new

special tokens to the models, named "[USER]" and "[CNSLR]." Those were added at the beginning of each message, presenting the conversation structure in a processable format to the models. For hyperparameter tuning, a grid search across the learning rate (2×10^{-5}, 3×10^{-5}, and 5×10^{-5}) and the batch size (1, 2, and 4) was performed for the preselected most promising model. The training and tuning were done at a stratified train-validation split (70:30 of the data used for algorithm training), as the repeated cross-validation principle applied for the TF-IDF approach was infeasible due to computational costs. Therefore, a train-validation-test split (56:24:20) was used as an evaluation principle, with the same data being kept aside as final test data for the nontransformer approach. All transformer-based models were trained on an NVIDIA GeForce RTX 3090 graphics processing unit with 24 GB video random access memory.

Final Evaluation

The 522 newest conversations with feedback were used as a test set. The distribution of the outcome did not differ significantly between the training and test data (t_{520}=–1.1; *P*=.30). We decided to predict the outcome with the best performing TF-IDF approach and the most promising transformer approach, as identified on the train set as described above. We then applied a permutation test [46] to evaluate the significance of both algorithms. Finally, we contrasted the achieved AUCs of the two approaches, applying a DeLong test [47], which has been suggested for this scenario [48]. We decided for this procedure above the 5×2 McNemar test [49] originally proposed in our preregistration. This reconsideration was mainly made due to the inability of the McNemar test to statistically compare AUC scores. The comparison of accuracies seemed disadvantageous to us, as focusing on the performance for one specific threshold. In contrast, considering the different proposed use cases, we were more interested in a threshold-independent comparison of classifier performance. As a threshold-dependent metric, we reported the Matthews correlation coefficient (MCC), which is particularly helpful in cases of imbalanced classes [50]. We followed the suggestion in the literature to use a default threshold of 0.5 [51] for the calculation of a confusion matrix and the corresponding MCC score.

Explainability

We used Shapley additive explanation (SHAP) values [52] as an explainability framework. This game-theory–based approach is applicable for transformer models [53] and XGBoost classifier [54].

Results

Algorithm Training

For the TF-IDF-based approach, the best set of hyperparameters selected through the tuning approach led to a mean ROC AUC score of 0.70 (SD 0.02) across repeated cross-validation for the XGBoost classifier. For this, a minimum occurrence of the word stems for 20 different chatters and for five different counselors was selected as a hyperparameter for the vectorizers. Random oversampling was selected for handling class imbalance. Counselors word stems were only selected when occurring in 30% or less of the conversations, while chatters word stems were allowed in up to 90% of the conversations. In addition, trigrams and bigrams were included, as well as predictors (see Table 2 for all hyperparameters). This was slightly above the performance of logistic regression, scoring 0.66 for the best set of hyperparameters.

Table 2. Overview of tuned hyperparameters (definitions adapted from [22]).

Hyperparameters	Description	Value range	Selected parameter
max_df_chatter	Terms that appear in more chatter documents than the threshold value are ignored. The value represents the proportion of documents	0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0	0.9
min_df_chatter	Terms that appear in fewer chatter documents than the threshold value are ignored	1, 2, 5, 10, 25, 50, 75, 100, 150, 200	20
max_df_couns	Analogous to max_df_chatterfor counselor messages	0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0	0.3
min_df_couns	Analogous to min_df_chatter for counselor messages	1, 2, 5, 10, 25, 50, 75, 100, 150, 200	5
Sampling method	Method for handling imbalance	ROS[a], RUS[b], SMOTE[c]	RandomOverSampler
colsample_bytree	Subsample ratio of columns for growing trees	0.2, 0.4, 0.6, 0.7, 0.8, 0.9, 1.0	1.0
eta	Learning rate	0.005, 0.01, 0.05, 0.1, 0.2	0.1
gamma	Minimum loss reduction to make a further split on a leaf node	0, 0.25, 0.5, 1, 1.5, 2, 5, 10	1.5
max_depth	Maximum depth of a tree	2, 4, 6, 8, 10, 12, 14, 16	16
min_child_weight	Minimum sum of instance weight (Hessian) needed in a child	1, 5, 10, 20	10
subsample	Subsample ratio of the training instances prior to growing trees	0.2, 0.4, 0.6, 0.7, 0.8, 0.9, 1.0	0.9
use_idf	Whether to term frequencies should be reweighted by the inverse document frequencies	True, false	True
ngram_range	Length of word sequences used as predictors	(1,1), (1, 2), (1,3)	(1,3)

[a]ROS: random over sampler.

[b]RUS: random under sampler.

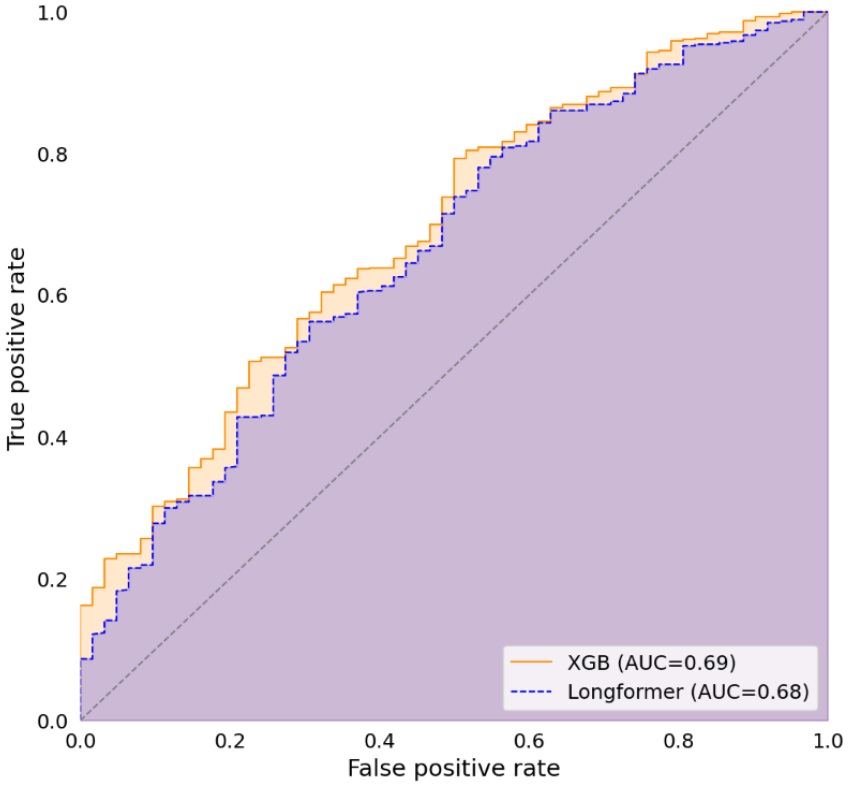
[c]SMOTE: Synthetic Minority Oversampling Technique.

For the transformer-based approach, we reached a ROC AUC of 0.58 for the DistilBERT and 0.59 for the GottBERT models, using class weights (9:1) and five epochs. Comparable performances were reached when random oversampling was used instead of the class weights. We expected the performance to be limited by strong truncation. Therefore, we explored the average length of the input sequence with DistilBERT as tokenizer. Data points in the train set contained on average 1889 (SD 873) tokens, showing that those models could just use a share of the available data on the chat conversations. However, with the longest conversation holding 8507 tokens, the Longformer model structure seemed capable of capturing nearly all information contained in our data. Indeed, using the Longformer model in combination with class weights (9:1), three epochs, a learning rate of 3e-5, and a batch size of one resulted in a significantly higher ROC AUC of 0.69. Neither other methods for handling class imbalance nor different epoch sizes lead to a further improved performance.

Final Evaluation

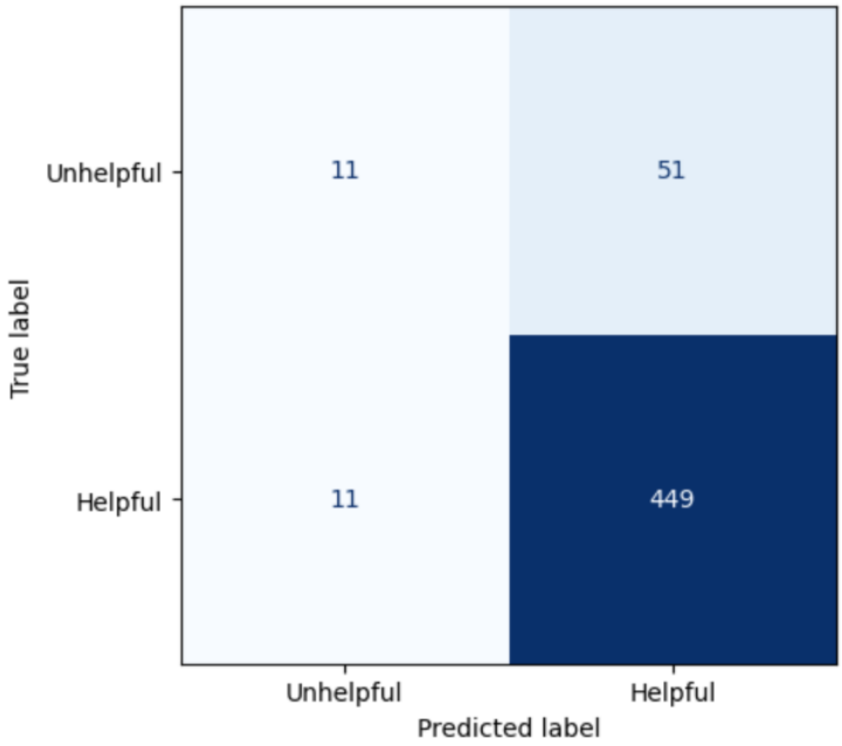
While the performance between the transformer and non–transformer-based approach was similar during training (0.69 vs 0.70), this comparison is limited by the differences in the used validation principle. However, the large previously unseen test set allowed us the comparison of the two best-of-class models in a final evaluation. Here, we reached an ROC AUC of 0.68 for the Longformer model and an ROC AUC of 0.69 for the TF-IDF–based approach, both significantly outperforming randomness in a permutation test (P<.001 for both). However, as expected, considering the similar performance, there was no significant difference between the two approaches (P=.69). The ROC curves are plotted in Figure 1, showing how threshold and model performance interacted.

Figure 1. ROC AUC curves comparing the two algorithms. AUC: area under the curve; ROC: receiver operating characteristic; XGB: extreme gradient boosting.



Consequently, we used the TF-IDF approach as the simpler algorithm for further insights, as well as the explainability approach. The average precision score here was 0.93 (SD 0.02) on the test set. The MCC score for the default threshold of 0.5 was 0.25 on the test set. The confusion matrix on this threshold can be found in Figure 2. Here, a positive predictive value of 0.90 and a negative predictive value (NPP) of 0.50 were achieved, with "positive" being coded as helpful. The sensitivity was 0.98 and the specificity was 0.18.

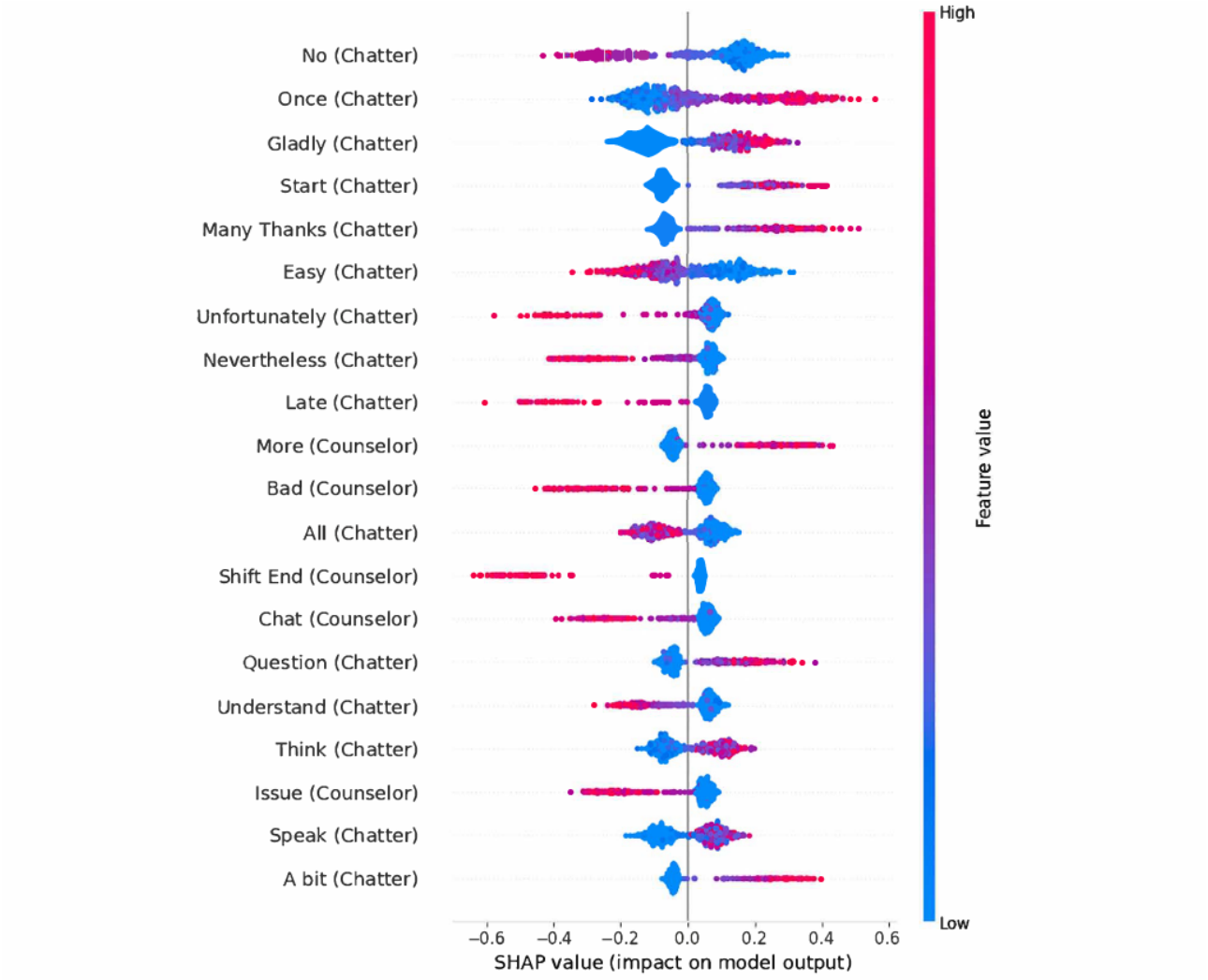
Figure 2. Confusion matrix for the selected threshold for the TF-IDF algorithm. TF-IDF: term frequency-inverse document frequency.



Explainability

We applied SHAP values on the vectorizer-based approach. The most predictive word identified here was “no” by the chatters, being associated with a higher chance of an unhelpful perceived chat. Two other predictors of unhelpfulness were the word “bad” (original: “schlimm”) by the counselor, as well as “nevertheless” (original: “trotzdem”) by the chatter, and “further on” (original: “weiterhin”) by the counselor. In addition, some bigrams were among the most predictive variables. For example, “shift end” (German: “Schicht endet”), indicating that a counselor had to end a conversation due to their shift being over, was associated with negative feedback. For an improved understanding of the context those words were used, we looked into chats using those and giving negative feedback afterward. While “no” was used in diverse settings, there was a notable number of cases where chatters denied the counselor’s offering of further help such as an exercise. “Bad” was used on several occasions where chatters reported highly traumatic experiences they had. Finally, “further on” was a phrase repeatedly used by counselors to announce the end of their shift and offer further support from a colleague afterward. There were also several words being predictive of perceived helpfulness. Several of those implied that a chatter expressed satisfaction with the interaction at the end of a chat. For example, the word stem “thanks” (original: “dank”) was predictive of higher perceived helpfulness, as was “great” (original: “toll”). We also investigated those conversations that were predicted with the highest likelihood of being labeled as unhelpful afterward. Again, there were several cases included where chatters rejected suggested exercises by the counselor. In addition, in several conversations with a high risk of unhelpfulness, it was reported that mental health care is already received, such as regularly seeing a psychiatrist or being hospitalized in a clinic. As one of the core functions of chat hotlines is the redirection into care, it might be harder to make a satisfying offer to those. The 20 most predictive words as identified by the tree-based SHAP approach can be found in Figure 3.

Figure 3. The 20 most predictive word stems as identified by the SHAP approach for the TF-IDF algorithm. SHAP: Shapley additive explanations; TF-IDF: term frequency-inverse document frequency.



Discussion

Primary Findings

This project investigated the use of NLP techniques for an automated evaluation of the perceived helpfulness of chat-based counseling. We were able to reach a ROC AUC of 0.67 on the previously unseen test set for a transformer, as well as for a non–transformer-based approach. Our explainability part revealed several linguistic markers of perceived unhelpful chat consultations such as the written expression of thankfulness, or the extensive use of the word "no" for rejecting the different offers made by counselors.

The reached performance was moderate, though significant and in line with past work from the identical settings [22]. However, the feasibility of an AI use case always depends on the performance considering the proposed use case. The given study implied two potential uses of predicted helpfulness of the chats.

The first use case was the real-time identification of unsuccessful consultations, as perceived by the chatter. Due to the very harmful impact of such experiences, those predictions could be used for a tailored follow-up, for example, with details of different treatment options for those affected. In our example, we would have identified 30 of the 62 unhelpful rated conversations with the approach, though 79% of all identified cases would have been false negatives (with negative referring to perceived unhelpfulness).

An alternative approach would have been a much stricter threshold, letting us mark significantly less chats but with higher NPP. For example, on a threshold of 0.3, our NPP would have doubled. However, the consequences of wrongly identifying chatters as unsatisfied might be less relevant than missing those being unsatisfied in light of the possible negative consequences of further help seeking. Overall, whether one of those approaches could be valuable would depend on whether the benefits for those correctly identified are larger than the costs of providing the intervention based on the prediction. Finally, this is an empirical question that we cannot answer here sufficiently. This highlights the large need for randomized controlled trials for prediction studies, moving from feasibility to actually showing clinical benefits [55].

A second use case of the proposed algorithm lies less on the individual and more on a population-based level. As evaluation within naturalistic and low-threshold settings is commonly difficult, the developed algorithm could be applied to those who did not respond to feedback questionnaires. This application would allow a better-informed estimation of satisfaction with the service where just a minority provides active feedback. A reliable estimate of this core metric of the service would propose a huge value for organizational purposes. Without any alternative of estimating the satisfaction of those not providing feedback being available, the proposed algorithm already provides an improvement over the status quo as clearly performing above the chance level. However, particularly for systematic comparison of, for example, monthly satisfaction, the question arises whether the performance is sufficient for reliable inference. Here, simulation studies might help to better understand the relation between performance and the reliability of algorithm-based evaluation.

Secondary Findings

Interestingly, there was no further gain in predictive capability by using the computational heavy and pretrained Longformer model. The failure of more complex NLP models to outperform simpler ones is not unique to the given setting and has been reported before [56-58]. However, based on the literature, we started the work on this paper with an opposing hypothesis. For example, a popular study [59] compared Bidirectional Encoder Representations from Transformer–based approaches with TF-IDF–based algorithms and reported a clearly better performance for the former. An in-depth look into the used methods provides several possible explanations for the diverging results. First, the cited study used a larger sample of 50,000 distinct cases, while using the much smaller Bidirectional Encoder Representations from Transformer base model. Therefore, the dataset size might have been insufficient to finetune such a sophisticated model. Second, the use case is different, while algorithmic performance is highly case specific. The cited study focuses on sentiment analysis. Arguably, the extraction from word vectors into higher-dimensional spaces like sentiment as done by transformer models is particularly relevant here. While our explainability approach revealed some sentiment-related predictors like words of thankfulness, overly sentiment seemed less central than it is for movie reviews as in the aforementioned study. Finally, it remains unclear how much the advantage of simpler models is used in comparative studies. For example, in our approach, we were able to perform extensive hyperparameter tuning using sophisticated cross-validation principles. The relevance of this to produce generalizable results, and therefore, realistic performance estimates is well established [60,61]. Such approaches are hard to reproduce at feasible computational costs for transformer-based models for a lot of ML practitioners in their day-to-day work. However, waiving those techniques also for the baseline is arguably biasing the comparison against them, as their better capability to be trained with extended cross-validation principles is a real benefit that might translate into predictive performance. Particularly, small predictive performance differences as reported regularly (eg, [25]) might disappear with decent hyperparameter tuning and cross-validation.

In conclusion, while the actual outperformance seems dependent on setting and data, the results of this study, as well as the aforementioned studies, highlight the relevance of benchmarking complex models with simpler ones. Otherwise, overly complex models might be implemented without benefits. There are numerous studies that apply interesting and promising algorithmic approaches but do not compare them with a simpler baseline at all (eg, [62-64]). However, we also argue that a fair comparison includes the utilization of hyperparameter tuning and cross-validation for computationally lighter models.

Limitations

There were limitations to the approach in this paper. First, while we predicted the helpfulness of a chat as perceived by chatters, this perception does not equal to actually being clinically beneficial. For example, in the aforementioned study by Imel

et al [27], the association between message content and satisfaction was much stronger than the association between content and symptom reduction. Therefore, future work could benefit from associating chat messages with clinically validated questionnaires as output. However, arguably changes in symptoms are difficult to measure in hotline settings, where a majority of chatters just contact the service once. Second, we were only able to train the algorithms on the data of those who responded to the feedback questionnaire. This might have introduced a bias, in case of systematic differences between those providing feedback and those who do not. Third, we focused on the application of the Longformer model in the transformer-based approach of this paper. Future work might also benefit from exploring task-specific adaptions of the used algorithms in detail. In addition, different methods of handling long text inputs such as BELT [65] might enable a better performance. Notably, there were no mental health–specific smaller models available in German. Those exist for other languages and use cases [66]. Such models, for example, pretrained on youth mental health data in German, could provide further performance gains as well. Finally, while we used a test set for a final one-time evaluation, this test set still came from the same chat counseling service. However, the relevance of truly external test sets has been highlighted repeatedly as being relevant for more valid claims regarding the generalizability of a chosen approach (eg, [67]).

Conclusions

In summary, there is a predictive signal regarding the perceived service quality in the chat messages at a 24/7 chat hotline for youth. This opens interesting use cases in the quality control and evaluation efforts at those hotlines. Future work such as the randomized evaluation of interventions based on the predicted helpfulness is needed for moving toward real-world implementation.

Acknowledgments

We acknowledge support by the Open Access Publication Fund of Humboldt-Universität zu Berlin.

Authors' Contributions

SH developed the idea, analyzed the data, and wrote the first draft of the paper. All authors contributed to the development of the exact analysis to be performed. All authors reviewed and contributed to the final draft.

Conflicts of Interest

SH and RW are employed by krisenchat, the organization that provided the data for this study. SH is also employed by Elona Health, a provider of digital health applications for mental health in Germany. KH is a scientific advisor and received virtual stock options from Mental Tech GmbH, which develops an artificial intelligence–based chatbot providing mental health support.

Multimedia Appendix 1
Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies.
[DOCX File , 20 KB - ai_v4i1e63701_app1.docx]

Multimedia Appendix 2
Full questionnaire sent out to chatters, original (German) and English translation.
[DOCX File , 16 KB - ai_v4i1e63701_app2.docx]

References

1. Kessler RC, Angermeyer M, Anthony JC, de Graaf R, Demyttenaere K, Gasquet I, et al. Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey initiative. World Psychiatry 2007;6(3):168-176 [FREE Full text] [Medline: 18188442]
2. de Girolamo G, Dagani J, Purcell R, Cocchi A, McGorry PD. Age of onset of mental disorders and use of mental health services: needs, opportunities and obstacles—CORRIGENDUM. Epidemiol Psychiatr Sci 2022;31:e46 [FREE Full text] [doi: 10.1017/S2045796022000282] [Medline: 35762753]
3. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. Lancet Psychiatry 2016;3(2):171-178. [doi: 10.1016/S2215-0366(15)00505-2] [Medline: 26851330]
4. Christensen MK, Lim CCW, Saha S, Plana-Ripoll O, Cannon D, Presley F, et al. The cost of mental disorders: a systematic review. Epidemiol Psychiatr Sci 2020;29:e161 [FREE Full text] [doi: 10.1017/S204579602000075X] [Medline: 32807256]
5. McGorry PD, Mei C. Early intervention in youth mental health: progress and future directions. Evidence Based Mental Health 2018;21(4):182-184 [FREE Full text] [doi: 10.1136/ebmental-2018-300060] [Medline: 30352884]
6. Colizzi M, Lasalvia A, Ruggeri M. Prevention and early intervention in youth mental health: is it time for a multidisciplinary and trans-diagnostic model for care? Int J Mental Health Syst 2020;14:23 [FREE Full text] [doi: 10.1186/s13033-020-00356-9] [Medline: 32226481]

7. Catania LS, Hetrick SE, Newman LK, Purcell R. Prevention and early intervention for mental health problems in 0–25 year olds: are there evidence-based models of care? Adv Mental Health 2014;10(1):6-19. [doi: 10.5172/jamh.2011.10.1.6]
8. McGorry PD, Mei C, Chanen A, Hodges C, Alvarez-Jimenez M, Killackey E. Designing and scaling up integrated youth mental health care. World Psychiatry 2022;21(1):61-76 [FREE Full text] [doi: 10.1002/wps.20938] [Medline: 35015367]
9. Tibbs M, O'Reilly A, O'Reilly MD, Fitzgerald A. Online synchronous chat counselling for young people aged 12-25: a mixed methods systematic review protocol. BMJ Open 2022;12(4):e061084 [FREE Full text] [doi: 10.1136/bmjopen-2022-061084] [Medline: 35470202]
10. Ersahin Z, Hanley T. Using text-based synchronous chat to offer therapeutic support to students: a systematic review of the research literature. Health Educ J 2017;76(5):531-543. [doi: 10.1177/0017896917704675]
11. Mathieu SL, Uddin R, Brady M, Batchelor S, Ross V, Spence SH, et al. Systematic review: the state of research into youth helplines. J Am Acad Child Adolesc Psychiatry 2021;60(10):1190-1233. [doi: 10.1016/j.jaac.2020.12.028] [Medline: 33383161]
12. Teens, social media and technology 2023. Pew Research Center. 2023. URL: https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/ [accessed 2024-01-30]
13. Hajok D. Der veränderte Medienumgang Jugendlicher. Tendenzen aus 20 Jahren JIM-Studie. The changing media usage of adolescents: trends from 20 years of the JIM study. Jugend Medien Schutz-Report 2018;41(6):4-6. [doi: 10.5771/0170-5067-2018-6-4]
14. Eckert M, Efe Z, Guenthner L, Baldofski S, Kuehne K, Wundrack R, et al. Acceptability and feasibility of a messenger-based psychological chat counselling service for children and young adults ("krisenchat"): a cross-sectional study. Internet Interventions 2022;27:100508 [FREE Full text] [doi: 10.1016/j.invent.2022.100508] [Medline: 35242589]
15. Thompson LK, Sugg MM, Runkle JR. Adolescents in crisis: a geographic exploration of help-seeking behavior using data from Crisis Text Line. Soc Sci Med 2018;215:69-79. [doi: 10.1016/j.socscimed.2018.08.025] [Medline: 30216891]
16. Watling D, Batchelor S, Collyer B, Mathieu S, Ross V, Spence SH, et al. Help-seeking from a national youth helpline in Australia: an analysis of kids helpline contacts. Int J Environ Res Public Health 2021;18(11):6024 [FREE Full text] [doi: 10.3390/ijerph18116024] [Medline: 34205148]
17. Gould MS, Pisani A, Gallo C, Ertefaie A, Harrington D, Kelberman C, et al. Crisis text-line interventions: evaluation of texters' perceptions of effectiveness. Suicide Life Threat Behav 2022;52(3):583-595 [FREE Full text] [doi: 10.1111/sltb.12873] [Medline: 35599358]
18. Lee EE, Torous J, de Choudhury M, Depp CA, Graham SA, Kim HC, et al. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. Biol Psychiatry Cogn Neurosci Neuroimaging 2021;6(9):856-864 [FREE Full text] [doi: 10.1016/j.bpsc.2021.02.001] [Medline: 33571718]
19. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol 2018;14:91-118. [doi: 10.1146/annurev-clinpsy-032816-045037] [Medline: 29401044]
20. Hornstein S, Zantvoort K, Lueken U, Funk B, Hilbert K. Personalization strategies in digital mental health interventions: a systematic review and conceptual framework for depressive symptoms. Front Digital Health 2023;5:1170002 [FREE Full text] [doi: 10.3389/fdgth.2023.1170002] [Medline: 37283721]
21. Swaminathan A, López I, Mar RAG, Heist T, McClintock T, Caoili K, et al. Natural language processing system for rapid detection and intervention of mental health crisis chat messages. NPJ Digital Med 2023;6(1):213 [FREE Full text] [doi: 10.1038/s41746-023-00951-3] [Medline: 37990134]
22. Hornstein S, Scharfenberger J, Lueken U, Wundrack R, Hilbert K. Predicting recurrent chat contact in a psychological intervention for the youth using natural language processing. NPJ Digital Med 2024;7(1):132 [FREE Full text] [doi: 10.1038/s41746-024-01121-9] [Medline: 38762694]
23. Xu Y, Chan CS, Tsang C, Cheung F, Chan E, Fung J, et al. Detecting premature departure in online text-based counseling using logic-based pattern matching. Internet Interventions 2021;26:100486 [FREE Full text] [doi: 10.1016/j.invent.2021.100486] [Medline: 34877263]
24. Kohls E, Guenthner L, Baldofski S, Eckert M, Efe Z, Kuehne K, et al. Suicidal ideation among children and young adults in a 24/7 messenger-based psychological chat counseling service. Front Psychiatry 2022;13:862298 [FREE Full text] [doi: 10.3389/fpsyt.2022.862298] [Medline: 35418889]
25. Broadbent M, Grespan MM, Axford K, Zhang X, Srikumar V, Kious B, et al. A machine learning approach to identifying suicide risk among text-based crisis counseling encounters. Front Psychiatry 2023;14:1110527 [FREE Full text] [doi: 10.3389/fpsyt.2023.1110527] [Medline: 37032952]
26. Xu Z, Xu Y, Cheung F, Cheng M, Lung D, Law YW, et al. Detecting suicide risk using knowledge-aware natural language processing and counseling service data. Soc Sci Med 2021;283:114176. [doi: 10.1016/j.socscimed.2021.114176] [Medline: 34214846]
27. Imel ZE, Tanana MJ, Soma CS, Hull TD, Pace BT, Stanco SC, et al. Mental health counseling from conversational content with transformer-based machine learning. JAMA Netw Open 2024;7(1):e2352590 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.52590] [Medline: 38252437]
28. Li A, Ma J, Ma L, Fang P, He H, Lan Z. Towards automated real-time evaluation in text-based counseling. ArXiv. Preprint posted online on March 07, 2022 2022 [FREE Full text]

29. Rickwood D, Deane FP, Wilson CJ, Ciarrochi J. Young people's help-seeking for mental health problems. Aust e-J Adv Mental Health 2014;4(3):218-251. [doi: 10.5172/jamh.4.3.218]
30. de Winter AF, Oldehinkel AJ, Veenstra R, Brunnekreef JA, Verhulst FC, Ormel J. Evaluation of non-response bias in mental health determinants and outcomes in a large sample of pre-adolescents. Eur J Epidemiol 2005;20(2):173-181 [FREE Full text] [doi: 10.1007/s10654-004-4948-6] [Medline: 15792285]
31. Cheung KL, Ten Klooster PM, Smit C, de Vries H, Pieterse ME. The impact of non-response bias due to sampling in public health studies: a comparison of voluntary versus mandatory recruitment in a Dutch national survey on adolescent health. BMC Public Health 2017;17(1):276 [FREE Full text] [doi: 10.1186/s12889-017-4189-8] [Medline: 28330465]
32. Automated evaluation of helpfulness of chat-counseling sessions for the youth. a natural language processing study. OSF Registries. URL: https://osf.io/sr4q9 [accessed 2024-06-26]
33. Klement W, El Emam K. Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: development and validation. J Med Internet Res 2023;25:e48763 [FREE Full text] [doi: 10.2196/48763] [Medline: 37651179]
34. silvanhornstein/AutoEval: code for paper (OSF: SR4Q9). GitHub. URL: https://github.com/silvanhornstein/AutoEval [accessed 2024-06-26]
35. Hornstein S, Forman-Hoffman V, Nazander A, Ranta K, Hilbert K. Predicting therapy outcome in a digital mental health intervention for depression and anxiety: a machine learning approach. Digital Health 2021;7:20552076211060659 [FREE Full text] [doi: 10.1177/20552076211060659] [Medline: 34868624]
36. Wartena C. A probabilistic morphology model for German lemmatization. 2019. URL: https://serwiss.bib.hs-hannover.de/frontdoor/index/index/docId/1527 [accessed 2019-01-01]
37. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
38. Halimu C, Kasem A, Newaz S. Empirical comparison of area under ROC curve (AUC) and mathew correlation coefficient (MCC) for evaluating machine learning algorithms on imbalanced datasets for binary classification. 2019 Presented at: Proceedings of the 3rd International Conference on Machine Learning and Soft Computing; January 25-28, 2019; Da Lat, Vietnam p. 1-6. [doi: 10.1145/3310986.3311023]
39. McDermott MBA, Zhang H, Hansen LH, Angelotti G, Gallifant J. A closer look at AUROC and AUPRC under class imbalance. ArXiv. Preprint posted online on January 11, 2024 2024. [doi: 10.48550/arXiv.2401.06091]
40. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16(1):321-357. [doi: 10.1613/jair.953]
41. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
42. The AI community building the future. Hugging Face. URL: https://huggingface.co/ [accessed 2024-04-05]
43. Scheible R, Thomczyk F, Tippmann P, Jaravine V, Boeker M. GottBERT: a pure German language model. ArXiv. Preprint posted online on December 03, 2020 2020 [FREE Full text] [doi: 10.48550/arXiv.2012.02110]
44. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv. Preprint posted online on October 2, 2019 2019 [FREE Full text]
45. Beltagy I, Peters M, Cohan A. Longformer: the long-document transformer. ArXiv. Preprint posted online on April, 10, 2020 2020 [FREE Full text]
46. Ojala M, Garriga GC. Permutation tests for studying classifier performance. 2009 Presented at: 2009 Ninth IEEE International Conference on Data Mining; December 06-09, 2009; Miami Beach, FL. [doi: 10.1109/icdm.2009.108]
47. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837-845. [Medline: 3203132]
48. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep 2024;14(1):6086 [FREE Full text] [doi: 10.1038/s41598-024-56706-x] [Medline: 38480847]
49. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998;10(7):1895-1923. [doi: 10.1162/089976698300017197] [Medline: 9744903]
50. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. URL: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html [accessed 2025-02-04]
51. Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min 2021;14(1):13 [FREE Full text] [doi: 10.1186/s13040-021-00244-z] [Medline: 33541410]
52. Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Min 2023;16(1):4 [FREE Full text] [doi: 10.1186/s13040-023-00322-4] [Medline: 36800973]

53. Kokalj E, Škrlj B, Lavrač N, Pollak S, Robnik-Šikonja M. BERT meets Shapley: extending SHAP explanations to transformer-based classifiers. 2021 Presented at: Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation; February 03, 2025; Hackashop p. 16-21 URL: https://aclanthology.org/2021.hackashop-1.3/
54. Li Z. Extracting spatial effects from machine learning model using local interpretation method: an example of SHAP and XGBoost. Comput Environ Urban Syst 2022;96:101845. [doi: 10.1016/j.compenvurbsys.2022.101845]
55. Zhou Q, Chen ZH, Cao YH, Peng S. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. NPJ Digital Med 2021;4(1):154 [FREE Full text] [doi: 10.1038/s41746-021-00524-2] [Medline: 34711955]
56. Zantvoort K, Scharfenberger J, Boß L, Lehr D, Funk B. Finding the best match—a case study on the (text-)feature and model choice in digital mental health interventions. J Healthcare Inform Res 2023;7(4):447-479 [FREE Full text] [doi: 10.1007/s41666-023-00148-z] [Medline: 37927375]
57. Gogoulou E, Boman M, Abdesslem F, Isacsson N, Kaldo V, Sahlgren M. Predicting treatment outcome from patient texts:the case of internet-based cognitive behavioural therapy. 2021 Presented at: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume; February 03, 2025; Virtual event p. 575-580 URL: https://aclanthology.org/2021.eacl-main.46/ [doi: 10.18653/v1/2021.eacl-main.46]
58. Funk B, Sadeh-Sharvit S, Fitzsimmons-Craft EE, Trockel MT, Monterubio GE, Goel NJ, et al. A framework for applying natural language processing in digital health interventions. J Med Internet Res 2020;22(2):e13855 [FREE Full text] [doi: 10.2196/13855] [Medline: 32130118]
59. Garrido-Merchan EC, Gozalo-Brizuela R, Gonzalez-Carvajal S. Comparing BERT against traditional machine learning models in text classification. JCCE 2023;2(4):352-356. [doi: 10.47852/bonviewjcce3202838]
60. Bartz E, Zaefferer M, Mersmann O, Bartz-Beielstein T. Experimental investigation and evaluation of model-based hyperparameter optimization. ArXiv. Preprint posted online on July 19, 2021 2021 [FREE Full text]
61. Turner R, Eriksson D, McCourt M, Kiili J, Laaksonen E, Xu Z, et al. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: analysis of the black-box optimization challenge 2020. PMLR 2020;133:3-26 [FREE Full text] [doi: 10.1007/978-1-4842-6579-6_4]
62. Liu Z, Peach RL, Lawrance EL, Noble A, Ungless MA, Barahona M. Listening to mental health crisis needs at scale: using natural language processing to understand and evaluate a mental health crisis text messaging service. Front Digital Health 2021;3:779091 [FREE Full text] [doi: 10.3389/fdgth.2021.779091] [Medline: 34939068]
63. El-Ramly M, Abu-Elyazid H, Mo?men Y, Alshaer G, Adib N, Eldeen KA. CairoDep: detecting depression in arabic posts using BERT transformers. : IEEE; 2021 Presented at: 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS); December 05-07, 2021; Cairo, Egypt. [doi: 10.1109/icicis52592.2021.9694178]
64. Wang S, Dang Y, Sun Z, Ding Y, Pathak J, Tao C, et al. An NLP approach to identify SDoH-related circumstance and suicide crisis from death investigation narratives. J Am Med Inform Assoc 2023;30(8):1408-1417 [FREE Full text] [doi: 10.1093/jamia/ocad068] [Medline: 37040620]
65. mim-solutions / bert_for_longer_texts. GitHub. URL: https://github.com/mim-solutions/bert_for_longer_texts [accessed 2024-08-26]
66. Vajre V, Naylor M, Kamath U, Shehu A. PsychBERT: a mental health language model for social media mental health behavioral analysis. 2021 Presented at: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 9-12, 2021; Houston, TX p. 1077-1082. [doi: 10.1109/bibm52615.2021.9669469]
67. Chekroud AM, Hawrilenko M, Loho H, Bondar J, Gueorguieva R, Hasan A, et al. Illusory generalizability of clinical prediction models. Science 2024;383(6679):164-167. [doi: 10.1126/science.adg8538] [Medline: 38207039]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
MCC: Matthews correlation coefficient
NLP: natural language processing
NPP: negative predictive value
ROC: receiver operating characteristic
SHAP: Shapley additive explanations
TF-IDF: term frequency-inverse document frequency
XGBoost: extreme gradient boosting

Edited by K El Emam, B Malin; submitted 27.06.24; peer-reviewed by R Scheible, A Li; comments to author 17.08.24; revised version received 04.09.24; accepted 02.12.24; published 18.02.25.

Please cite as:
Hornstein S, Lueken U, Wundrack R, Hilbert K
Predicting Satisfaction With Chat-Counseling at a 24/7 Chat Hotline for the Youth: Natural Language Processing Study
JMIR AI 2025;4:e63701
URL: https://ai.jmir.org/2025/1/e63701
doi:10.2196/63701
PMID:

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org