IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 March 2025, accepted 24 March 2025, date of publication 31 March 2025, date of current version 11 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3556245

RESEARCH ARTICLE

Effect of Explainable Artificial Intelligence on Trust of Mental Health Professionals in an AI-Based System for Suicide Prevention

ADONIAS CAETANO DE OLIVEIRA[1,2], JOÃO PEDRO CAVALCANTI AZEVEDO[3], LÍVIA RUBACK[4], RAYELE MOREIRA[1], SILMAR SILVA TEIXEIRA[1], AND ARIEL SOARES TELES[1,3,5], (Member, IEEE)

[1]Programa de Pós-Graduação em Biotecnologia (PPGBiotec), Universidade Federal do Delta do Parnaíba, Parnaíba 64202-020, Brazil
[2]Campus Tianguá, Instituto Federal do Ceará, Tianguá 62320-000, Brazil
[3]Programa de Pós-graduação em Ciência da Computação (PPGCC), Universidade Federal do Maranhão, São Luís 65080-805, Brazil
[4]Faculdade de Tecnologia (FT), Universidade Estadual de Campinas, Campinas 13083-970, Brazil
[5]Campus Araioses, Instituto Federal do Maranhão, Araioses 65570-000, Brazil

Corresponding author: Ariel Soares Teles (ariel.teles@ifma.edu.br)

The Article Processing Charge for publishing this research was funded by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) (ROR identifier: 00x0ma614).

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee for Research Involving Humans of the Parnaíba Delta Federal University under Application No. 5.787.194, and performed in line with the Resolution No. 466/2012 of the Brazilian National Health Council, aligning with the Declaration of Helsinki.

ABSTRACT Artificial Intelligence (AI)-based systems have been proposed to aid Mental Health Professionals (MHPs) in various tasks, including the prevention of suicide by identifying Suicidal Ideation (SI). However, these systems may lack transparency and thereby create mistrust among MHPs. Explainable Artificial Intelligence (XAI) methods can elucidate how features influence system predictions, aiding MHPs in understanding them. This exploratory study aims to investigate how MHPs' trust is influenced by AI explanations (educational intervention and XAI methods) and other factors (professional background, knowledge of AI and computing, and reported system misclassification). We conducted an experiment using *Boamente*, an AI-powered clinical decision support system designed to assist MHPs in suicide prevention. *Boamente* identifies SI in Brazilian Portuguese texts typed on smartphones by leveraging a Large Language Model (LLM) for analysis. The results demonstrate that professional background, knowledge of AI and computing, and educational intervention had no statistically significant effect on trust. In contrast, trust was affected by factors such as LLM prediction explanations, the quality of explanations, and reported misclassification. Therefore, providing prediction explanations to understand the inner workings of AI models led MHPs to be more critical in relation to predictions, while there was an overtrust on MHPs when no explanations were provided. Furthermore, disagreement with LLM classifications and perceptions of system vulnerabilities also affected trust.

INDEX TERMS Explainable artificial intelligence, trust, medical AI, mental health, suicide, digital phenotyping.

I. INTRODUCTION

Suicide is a leading cause of premature mortality, potentially preventable [80]. The World Health Organization (WHO) reports that more than 700,000 individuals die by suicide each year, that is, one death every 40 seconds [113]. Also, around 77% of suicides in the world occur in low- and middle-income countries. This phenomenon impacts individuals across all regions and age groups worldwide [8], [29], [81]. Suicide is influenced by a complex interplay of factors such as

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

media coverage, economic conditions, biological influences, psychology issues, environmental factors, history context, and social dynamics [83], [90], [108]. Individuals at high risk for suicide often struggle with mental health disorders, such as bipolar disorder, anxiety, and depression [18], [73], [88]. The persistent experience of extreme hopelessness, helplessness, worthlessness, or feelings of defeat and confinement gradually drives individuals toward suicidal behaviour [86], [90], [108].

Suicidal behaviour includes a spectrum of actions ranging from Suicidal Ideation (SI) and planning to attempts and completed suicides [13]. SI, which precedes actual suicide, encompasses thoughts, desires, and preoccupations related to death and self-harm. This includes planning, writing farewell notes, and even expressing these feelings through social media [42], [74]. SI is associated with a variety of mental health disorders, such as alcohol use disorder, panic symptoms, anxiety, and sleep disturbances, as well as situational stressors [13], [42]. Among older adults, cultural acceptance of suicide, religiosity, and intergenerational cohabitation were predictors of SI [13]. Social media has increasingly become a central platform for people to express themselves and interact with others [58] For vulnerable groups, such as those struggling with mental health issues or substance abuse problems, these platforms may be used to discuss and acquire information about suicide [93].

Effective suicide prevention requires the active involvement of Mental Health Professionals (MHPs), such as psychologists and psychiatrists [62]. These practitioners are crucial to the prevention process, playing a central role in the identification of SI among their patients. Traditionally, SI analysis involves psychological evaluations, which include direct patient-MHP interactions and suicide risk identification based on self-reports and face-to-face interviews [38], [56], [98]. However, these approaches depend heavily on in-person interactions, requiring extensive training, ongoing supervision, workflow adjustments, and considerable time and effort, which limits their scalability [16].

Digital interventions based on Artificial Intelligence (AI) are developed to improve mental health support. These novel approaches include clinical decision support systems like *Boamente* [31], [32], which was proposed to detect and monitor SI remotely through patients' mobile devices (e.g., smartphones and tablets). The *Boamente* system passively collects non-clinical texts written in Brazilian Portuguese (PT-BR) from patients through a customized virtual keyboard application. These texts are transmitted to a web platform, where an AI model classifies them as indicative of negative or positive for SI (i.e., a binary classification task). The web application then presents these classifications on dashboards with graphical representations, allowing MHPs to monitor remotely patients' SI.

AI has been proposed as a transformative tool of mental health care, particularly as a decision-making support system [10], [20]. However, to effectively use AI-based decision support tools, MHPs must trust them. This trust is essential for successfully integrating AI as a complementary healthcare strategy [6], [39]. Lack of transparency, due to complex algorithms and misinterpreted data, can erode trust and disrupt patient-MHP relationships [3], [47], [85].

Although the *Boamente* system [32] shows promise, it uses a Large Language Model (LLM) called BERTimbau-Large [99], an opaque AI model (i.e., it is very difficult to understand the internal process that led to the output) [57], [99], which exacerbates trust issues by not explaining how it reaches its conclusions. This lack of transparency raises concerns about accountability and responsibility, which are important factors for building trust in AI systems [2], [6], [39]. Providing explanations for *Boamente* predictions has the potential to help MHPs better understand and trust the tool [2], [3]. The field of Explainable Artificial Intelligence (XAI) provides methods to clarify how SI-related content in texts affects *Boamente* predictions, i.e., explanations seek to help justify system predictions [2], [6].

This exploratory study addresses the trust issue of MHPs in the *Boamente* system for detecting SI. We evaluated AI explanations (educational intervention and XAI methods) and other factors, such as professional background, knowledge of AI and computing, reported misclassification (i.e., any misclassification of sentences or inaccuracies in classification probabilities), on the trust of MHPs through four groups of participants. Specifically, we utilized the Local Interpretable Model-Agnostic Explanations (LIME) method [92] to clarify the BERTimbau-Large model's decisions. This study fills a gap in the literature by specifically examining the trust of MHPs in an AI system designed for SI detection [12], [61]. MHPs play a central role in the application of AI in mental health. Understanding their trust factors is essential for developing accurate AI systems accepted and utilized by them [11].

Building trust between humans and AI is a significant challenge in the field of AI [11], [116]. Our study contributes to the growing body of research on human-AI interaction, particularly in the mental health care context, and the development of future AI systems for mental health applications. Our investigation of the existing literature suggests that this is the first study to explore XAI and trust in the context of an AI system designed to identify SI [35], [94], [107].

The remainder of this paper is organized as follows. Section II introduces the concepts related to this study. In Section III, we review the relevant literature and highlight the study contributions. Section IV presents material and methods. Section V presents the results, while Section VI discusses the findings, strengths, and limitations of this study. Finally, Section VII contains the conclusions of this study.

II. BACKGROUND

A. DIGITAL MENTAL HEALTH

Mental disorders are characterized by disruptions that have a clinical impact on a person's thoughts, emotional regulation,

cognition, behaviour, social interactions, and biological patterns [9], [26], [101]. Although they do not always present obvious physical symptoms, this health condition can cause distress or inability in activities related to study, work, family, or social life of those affected [26], [91].

Social networks have become popular to share different types of content, such as posting personal status updates and uploading photos. Furthermore, users can also interact with others by commenting on their posts and establishing conversations [112]. Previously, individuals with suicidal tendencies had the only option to express their feelings through suicide notes, whose linguistic and content analysis became the subject of several studies [5]. With the growing popularity of social networks and messaging apps, internet users tend to share content related to sadness, disappointments with life and even their suicidal thoughts, feelings and behaviours on these platforms [112].

Faced with these situations, adopting digital mental health tools to support MHPs in diagnosis, monitoring, and interventions is a complementary solution [105], [106]. Digital mental health refers to digital services and the adoption of technologies to support mental health professionals [66], [106]. Digital Phenotyping of Mental Health (DPMH) involves using personal sensing to passively collect data from ubiquitous devices [65], [66], [105], and then identify and monitor people with mental health disorders [65]. DPMH relies on mobile devices such as smartphones and wearables to analyse diverse social and behavioural data [66], [69]. DPMH solutions like *Boamente* can provide valuable insights into mental health care by continuously detecting patient-related information. For instance, data from smartphones and wearables can reveal details about a patient's environment, behaviour, and psychological states, including factors such as insomnia, motor activity, heart rate, sociability, and typed texts [66], [69], [71]. Leveraging technologies based on DPMH can empower healthcare professionals to make informed decisions and deliver personalized behavioural health interventions [64], [69].

Innovative and comprehensive solutions are crucial. Supporting the MHPs in preventing and delivering mental health interventions in the patients can make significant progress in enhancing public mental health [81], [95].

B. EXPLAINABLE ARTIFICIAL INTELLIGENCE

XAI focuses on developing methods for providing more transparent, interpretable, high-performing and reliable models [2], [3], [109]. It is common in the literature to use terms such as interpretable, explainable, transparency and comprehensibility interchangeably. However, each has a distinct definition [19], [89]. Transparency refers to the understandability of a model without insights into the actual algorithmic process. The understandability is associated with the complexity of the AI-based model and the ability of the algorithms to present their results in human terms. Explainability is defined as the details and reasons a model provides to make its operation clear and easy to understand for a given audience [89]. Interpretability is the ability of the AI-based system to be explained in human terms [89]. In other words, interpretable systems are those whose operations can be understood by humans [19].

Two main approaches to addressing transparency are glass box and closed box components [2], [33]. The term ''glass box'' in XAI refers to Machine Learning (ML) models that produce results easily understandable to experts in the application domain [57]. Glass-box or glass box components are transparent, and necessary information for understanding the model prediction is readily available [2], [33]. The term ''glass box'' might be misleading, as not all aspects of the model are always transparent. Therefore, some researchers use the term ''grey-box'' to indicate varying levels of transparency based on the available details [2], [60].

The term ''closed box'' is primarily used to describe ML models that are, from a mathematical perspective, very difficult for experts in practical domains to explain and understand [57]. Closed box models hide their internal mechanisms and logic from users and developers, making it challenging to fully understand the reasons behind their output. High-performance decision support systems often remain complex closed box entities [19], [33]. Lack of transparency creates dissonance because these systems do not offer clear explanations for their predictions [33].

XAI methods can be categorized into two main types: model-specific and model-agnostic. Model-specific methods provide explanations unique to a particular model or algorithm family [54], [67]. For instance, Linear Regression Models (LRMs) lack a direct measure of feature importance, but their coefficients can serve as an alternative explanation [67]. Model-agnostic methods can be applied to any algorithm [54] by analysing input-output pairs to explain model behaviour [67]. The flexibility of model-agnostic techniques makes them advantageous over model-specific ones [70]. Additionally, interpretability methods can be classified by scale: local methods provide specific explanations for each observation in an ML model, while global methods provide insights into the entire model [54]. For example, a global explanation may involve visualizing words from the model's training dataset that characterize positive and negative classes for SI. In contrast, a local explanation could focus on individual word importance scores in predicting one specific sentence.

Some ML models, such as LRMs and Decision Trees (DT), are intrinsically interpretable. Weights in LRM can indicate the importance of attributes, especially when the attributes are on a similar scale. For DTs, attribute importance can, for example, be calculated based on the average Gini impurity decrease [54], [63]. Conversely, one of the goals of XAI is to provide post-hoc methods for explanations after a model is trained [54]. Among the post-hoc methods, many are model-independent, employed to interpret any models and provide transparency [33]. While many of these methods are

model-independent, some only apply to a specific category of models, such as models based on neural networks.

Some XAI methods for explaining sentiment/emotion predictions in texts are LIME [92], Shapley Additive exPlanation (SHAP) [59], and ELI5 (''Explain Like I'm Five'') [34], [36]. Based on perturbations, LIME creates an explanation by approximating the complex ML model using an interpretable model, such as a linear model, trained on perturbations of a single input sample of interest [70]. LIME was employed in this study.

C. TRUST IN AI MEDICAL SYSTEMS

Trust is a human conviction or attitude [11], [50] that centres around the belief in the trustee's ability to meet the trustor's expectations [104], [111]. It is essential for effective healthcare systems, encompassing the professional-patient relationship and care tools. Trust is crucial for the adoption of AI by healthcare professionals and its broader integration into healthcare [4], [39].

Lack of trust (or distrust) can lead humans not to use an AI medical system, even if the system can benefit users (e.g., health professionals, patients) [52]. The ''optimal'' trust is not the maximum trust (or excessive trust) in AI systems because it does not lead to ideal decisions. Users with overtrust may accept results uncritically, which is risky in clinical settings. Overtrust can result in errors and negative impacts when a human user places more trust in an automated system than their actual abilities and limitations would allow [46], [48]. Optimal trust involves a certain level of mutual scepticism between users and AI-based systems. Because both can make mistakes, AI development must include mechanisms to maintain an optimal level of trust [11].

Despite their potential, many AI medical systems are underutilized or not used as intended [82]. This may be related to challenges, including the lack of transparency and issues with biased models [3], [47], [85]. Lack of transparency can lead to distrust and negatively impact the professional-patient relationship [3], [85]. Additionally, the complexity of the AI model makes it challenging for users to interpret predictions, which affects trust and acceptance [85].

In addition to explainability, transparency, and interpretability (see Section II-B), education can increase healthcare professionals' trust in medical AI [11], [107]. Understanding and trust are related and can, therefore, be affected by an educational intervention [52]. Educational intervention as part of AI literacy can be defined as educating users towards a better understanding of AI technology, including knowing and understanding AI [52], [55], [76]. Educational intervention based on learning AI's basic concepts through video may be very effective [48].

III. RELATED WORK

This section overviews relevant studies on (i) XAI in tools developed for suicide prevention, and (ii) the factors influencing trust in AI-based medical systems. We also compare our exploratory study to the existing body of knowledge to highlight its contributions.

A. XAI IN TOOLS FOR SUICIDE PREVENTION

Several studies propose AI models to assist in identifying manifestations related to mental health, such as suicide, including analysis in structured [77] and unstructured [41] data. Natural Language Processing (NLP) techniques examine SI manifestations in text data, such as interviews, social media posts and clinical notes [115]. According to Balcombe and De Leo [12], XAI should be adopted in implementing digital mental health for positive and responsible results.

Based on a plot showing the feature importance of predictors for suicide attempts and suicides, Nielsen et al. [77] elucidated that the risk of suicide initially increases with the number of attempts, but diminishes after 4 to 10 attempts. Additionally, the authors observed that diagnostic categories such as sociodemographic factors, mood disorders, and neurotic disorders influence suicide attempts and suicides in distinct ways. Tang et al. [102] used SHAP to rank feature importance in medical tabular data related to self-destructive behaviours and mental health issues, identifying key factors influencing suicide risk. The analysis revealed anger, depression, and social isolation as the primary predictors of suicide risk. In contrast, individuals with high income, respected professions, and higher education levels were found to have the lowest risk.

Malhotra et al. [61] evaluated the explainability of BERT-based LLMs to detect depressive and suicidal behaviour using SHAP and LIME methods. These XAI methods helped identify the causes of misclassification and provided insights into training data quality. SHAP values for these false positives helped the authors understand the words contributing to the misclassifications. Lekkas et al. [53] analysed the performance and interpretability of different ML models using SHAP to predict acute SI on public Instagram user data and post content. The SHAP results suggested that interaction data (e.g., followers, following, engagement, and similarity-based metrics) could be particularly useful predictors of future acute SI.

Nordin et al. [78] analysed the importance of features for suicide attempts provided by SHAP. The analyses revealed that the history of suicide attempts, SI, and ethnicity are predictors for suicide attempts. The explanatory results helped identify and clarify risk factors for suicide attempts, enhancing the understanding of suicide attempt prediction.

Previously, we started the process of explaining the *Boamente* system [31], focusing specifically on the ML models analysed in [32]. By using ELI5, we identified that Extra Trees and Random Forest were equally influenced by the features of one or two terms. The Support Vector Machine classifier was more influenced by features composed of two terms than just one. In general, traits including ''suicide'', ''desire to kill oneself'' and ''sadness'' had greater importance in indicating positive for SI.

B. FACTORS INFLUENCING TRUST IN AI-BASED MEDICAL SYSTEMS

Evans et al. [35] reviewed studies on the acceptability and trustworthiness of clinical decision support systems for health professionals. The authors found that transparency, explainability, and supporting evidence are key factors influencing healthcare practitioners' trust in AI-based systems. Rosenbacke et al. [94] reviewed how XAI can influence clinicians trust in AI applications in healthcare. The majority of studies suggest that XAI has the potential to enhance clinicians' trust in AI-generated recommendations, although some studies showed no effects or reached no conclusions. Here, we focus on studies that explore the factors influencing users' acceptance of AI-powered medical systems, with trust being a key factor in their adoption.

Zhang et al. [116] conducted an experimental study on the effects of AI explanations on understanding radiology reports. They found that providing model performance information can enhance trust and perceived usefulness, whereas providing local explanations improves understandability without necessarily increasing trust. Additionally, lower model performance tends to decrease trust with more disclosed information. Zhang et al. [116] also found that human trust in AI is influenced by the agreement with AI predictions and their correctness.

Tanguay-Sela et al. [103] assessed an AI-based clinical decision support system for depression treatment, examining its perceived utility, impact on clinical decision-making, and differences in usefulness between primary care providers and psychiatrists. The study revealed that familiarity with the tool significantly influences physician trust, with many participants preferring their own judgment over the system, relying more on their experience. Primary care providers found the model more helpful for treatment decisions compared to psychiatrists, likely due to the challenges they face in treating depression and their familiarity with relevant guidelines and medications.

Bourla et al. [15] examined the acceptability of new technologies for depression care among student nurses, focusing on Ecological Momentary Assessment (EMA), Computerized Adaptive Testing (CAT), and a Connected Wristband (CW). While CW was perceived as less useful than CAT and EMA, it was believed to detect depressive relapse more reliably, though EMA and CAT were not trusted for accurate diagnoses. Acceptability varied by age, sex, and curriculum semester, with younger students and women showing more negative perceptions. In a related study, Bourla et al. [14] assessed psychiatrists' acceptability of similar technologies, finding CW to have the lowest acceptability and CAT and EMA rated as more reliable but less usable. Acceptability was influenced by professional culture, with younger psychiatrists being more sceptical and about half desiring stronger scientific validation, suggesting that lack of knowledge, rather than outright rejection, shaped perceptions.

C. STUDY CONTRIBUTIONS

Similar to the first class of related works (Section III-A), the application domain of this exploratory study is the identification of SI. We applied the LIME method to explain predictions of the *Boamente* system using feature importance and examples-based explanations. In addition, we evaluated the effect of the XAI on the trust of MHPs. Our objective was not to uncover insights provided by XAI, such as listing predictors for SI, but rather to investigate how AI explanations influence MHPs' trust in the *Boamente*.

Related to the second category of works (Section III-B), this study seeks to address the *Boamente* acceptance by assessing MHPs' trust when providing an education intervention and explanations for classifications made by the BERTimbau-Large model. We collected the opinions from 78 participants (psychologists or psychiatrists) through questionnaires that assessed trust and the quality of explanations. We also investigated whether there was an effect of factors related to professional background, knowledge of AI and computing, reported misclassification, educational intervention, and AI explanations on the trust of MHPs.

IV. MATERIALS AND METHODS

In this section, we first describe the *Boamente* system, which was experimented by MHPs. Next, we explain the recruitment process, study design, data collection, and the analysed factors. Finally, we describe the statistical methods applied.

A. THE BOAMENTE SYSTEM

For practical purposes, the *Boamente* system was modified for a dedicated experiment, targeting only MHPs. This adaptation enabled participants to create and test sentences, facilitating the identification of misclassifications and gathering data on the AI system's reliability and effectiveness within a controlled, need-specific environment. The adapted version of *Boamente* also utilized the BERTimbau-Large model [99], a LLM that has shown promising performance across various tasks [100]. This model was trained on the original *Boamente* dataset [32] (refer to the Data Availability section for more details).

The *Boamente* interfaces were developed by combining Google Sites and Gradio tools [1] (refer to the Data Availability section for access to the code). Gradio is an open-source Python package that allows researchers to quickly generate visual interfaces for AI models. These interfaces were made available through a web application, which could be accessed via the research website's homepage. On this homepage, all participants could view a 5-minute video providing an overview of the research.[1] The website also provided information about the objectives and features of the *Boamente*, scientific articles of the project, contact emails of the researchers, and access to the Informed Consent Form (ICF). The *Boamente* interfaces could only be accessed

[1] Link to the video with audio in PT-BR: https://youtu.be/uhs-1zBvr_c

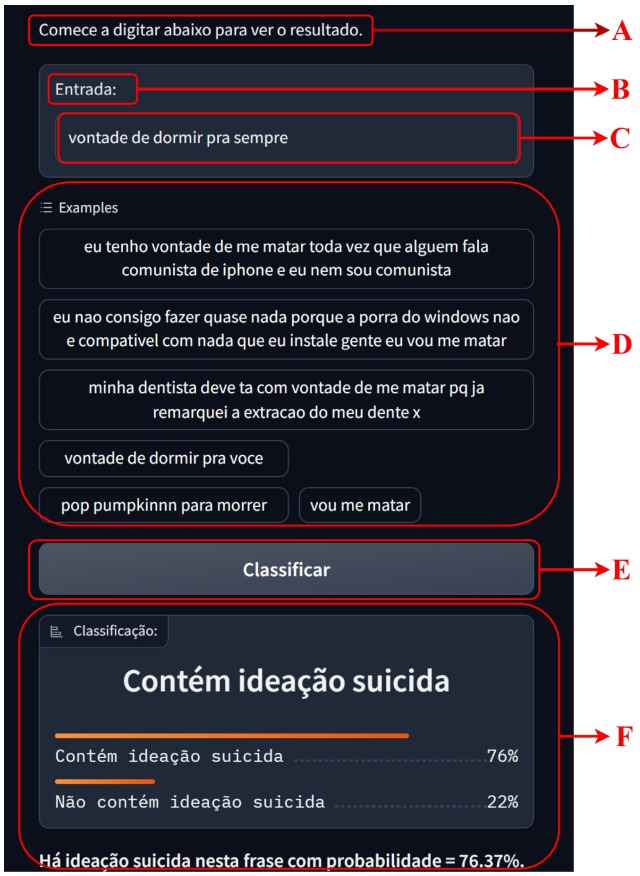


FIGURE 1. Screenshot of the interface integrated with the classification module.

on this website using a login and password, which were released only after participants were randomly assigned to their respective group.

Two interfaces were developed: an interface without explanations and another interface with explanations. Both interfaces contained the classification module. This module provides access to the text field to type sentences and the function to classify sentences. In this module, participants could test sentences related to SI and review their classification by the LLM. Figure 1 displays the interface integrated with the classification module, featuring the following graphical elements: (A) label with the text "Start typing below to see the result."; (B) label "Input"; (C) text field with an example of the sentence "I want to sleep forever" typed by the user; (D) examples of sentences randomly selected from the *Boamente* dataset, i.e., sentence examples were loaded with each run, allowing participants to test them; (E) classify button; and (F) classification result, indicating a 76% probability that the sentence contains SI and a 22% probability that it does not. Below, it states "There is suicidal ideation in this sentence with a probability = 76.37%".

Figure 2 displays the interface with explanations provided by the explanation module. This module provides access to the explanation function, which offers insights through five example sentences per class, word-level scores for the typed

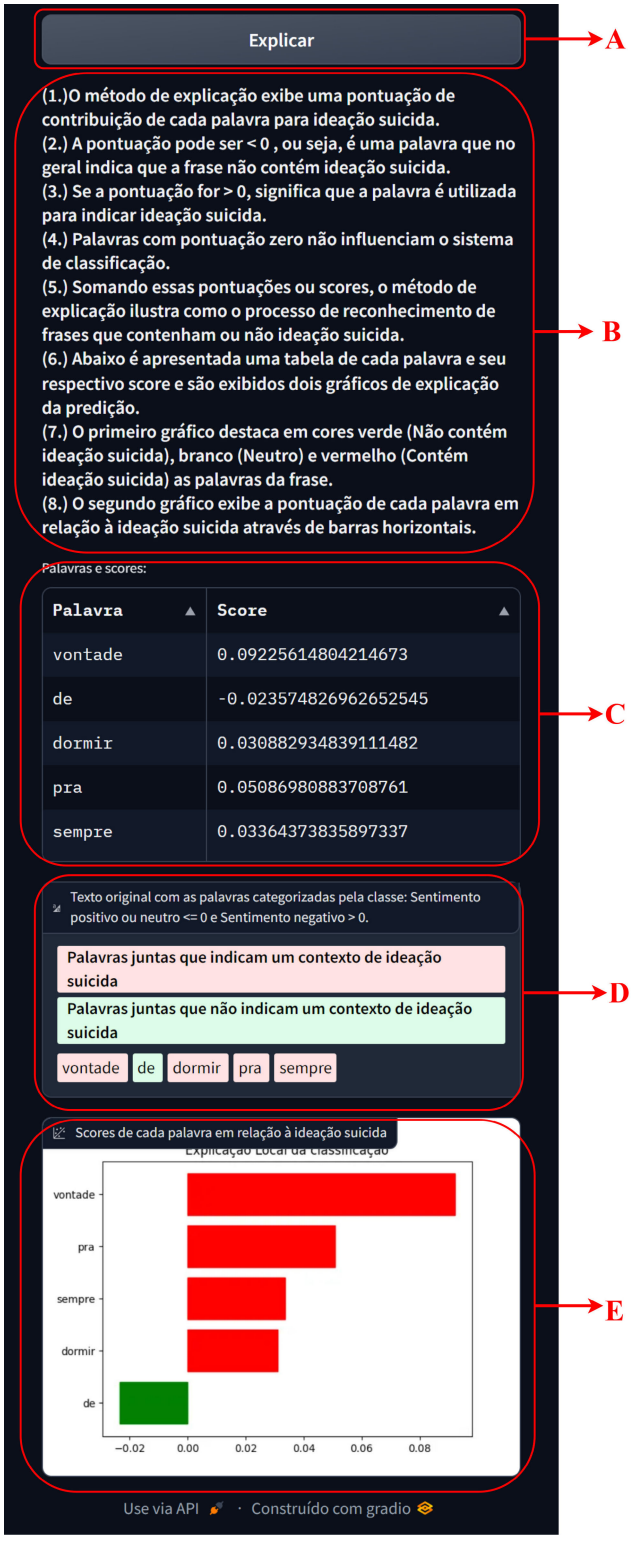


FIGURE 2. Screenshot of the interface integrated with the explanation module.

sentence, and an explanation based on word importance. The interface integrated with the explanation module contains the following graphical elements (Figure 2): (A) explain button; (B) guidelines on how to interpret explanations; (C) table of

scores of each word in the sentence; (D) LIME-based salience maps for text (the colour of each word corresponds to the class ''It contains suicidal ideation'' for red colour, and ''It does not contain suicidal ideation'' for green colour); and (E) chart of word importance.

B. PARTICIPANTS

Several recruitment strategies were employed during six months to ensure a diverse and representative sample of participants. These strategies included purposeful sampling, word of mouth, and snowball sampling. Purposeful sampling was used to select individuals with specific characteristics relevant to the research aims. Word of mouth involved participants and academic/university members sharing information about the study with potential candidates, while snowball sampling relied on referrals from initial participants to reach others who met the study criteria.

Once potential participants were identified, a research team member initiated contact via telephone or email. During this communication, the study, including its objectives, procedures, and expectations, was explained in detail. The team member also determined each individual's eligibility based on predefined selection criteria and assessed their interest and willingness to participate voluntarily. This approach ensured that participants were fully informed, with their involvement being both voluntary and based on their informed consent.

The sample consisted of MHPs, i.e., professionals in the mental health field with a psychology or psychiatric medicine degree from Ceará and Piauí, which are states in northeast Brazil. The inclusion criteria for recruiting participants were: (i) having a degree in psychology or medicine with a medical residency (i.e., postgraduate training program for physicians in Brazil) or specialization in psychiatry; (ii) working clinically in the therapy or treatment of patients. As exclusion criteria, we defined: (i) any condition that limits the participant's ability to evaluate the system; and (ii) refusal to sign the ICF. All participants received information about the study and signed the informed consent form. Ethical approval was obtained from the Ethics Committee for Research Involving Humans of the Parnaíba Delta Federal University (number 5.787.194).

Based on the survey conducted in September 2023, on the records of the Federal Councils of Medicine and Psychology, there are 3,353 psychologists and 466 psychiatrists in Ceará, while there are 10,550 psychologists and 152 psychiatrists in Piauí, resulting in a population of 14,521 MHPs [21], [22]. The minimum sample size was estimated based on a heterogeneous sample, with a 6% precision level, a 90% trust level, and a 10% margin of error. The estimated minimum sample size was at least 68 participants, averaging $n = 17$ participants per group. In total, 83 people were recruited for the research, but five did not complete all stages of the study. Therefore, the analysis of the results was limited to 78 participants who completed all stages of the research.

C. STUDY DESIGN AND PROCEDURE

This research was conducted mainly in an online format involving MHPs. We chose this online exploratory method because it expanded the sample's scope and diversity, offered greater flexibility for participants, and reduced costs and time [40]. The study design was inspired by the methodology adopted by Leichtmann et al. [52], and was organized into the following stages: (1) participants signed the ICF and completed the sociodemographic questionnaire (Table S1 - Supplementary Material), (2) took the AI knowledge test, (3) conducted experimental testing with the *Boamente* system, and (4) completed an evaluation questionnaire. Figure 3 illustrates these stages, which are detailed below.

All participants received explanations about the study's objectives, potential risks, and benefits and signed an ICF upon agreeing to participate. After that, they answered a sociodemographic questionnaire. After that, participants answered the AI knowledge test. Two conditions were considered:

- Educational intervention: we evaluated the effect of an educational intervention through a 20-minute online video explaining how *Boamente* works and the technologies it uses. This intervention approach was adopted because an adequate educational intervention should be more immersive and vivid than just text and images [48], [52]. Also, AI literacy allows users to explore the limits of an AI-based system [55];
- Explanations: as illustrated in Figure 4, we provided explanations for each sentence typed by the participant through dominant words, the importance of each word in the prediction, and the classification of text as positive or negative for SI.

Participants were then randomized into four groups, following a single-masked design, as illustrated in Figure 3. Each participant interacted with a *Boamente* interface corresponding to their assigned group, namely:

- Group A: accessed an interface without explanations and educational intervention;
- Group B: accessed an interface with educational intervention, but without explanations;
- Group C: accessed an interface with explanations but without educational intervention;
- Group D: accessed an interface with explanations and educational intervention.

In all groups, participants were instructed to experiment with *Boamente* by typing sentences in PT-BR that could or could not express SI. Each participant freely chose what to write, with no researcher intervention in selecting the types of sentences used by the MHPs. After typing sentences, they clicked on the ''classify'' button (Figure 1), and only group C and D participants could click on the ''explain'' button (Figure 2). This approach implies that participants in groups A and B performed sentence classification using a closed box model, while those in groups C and D had access to explanations for predictions, using an explainable

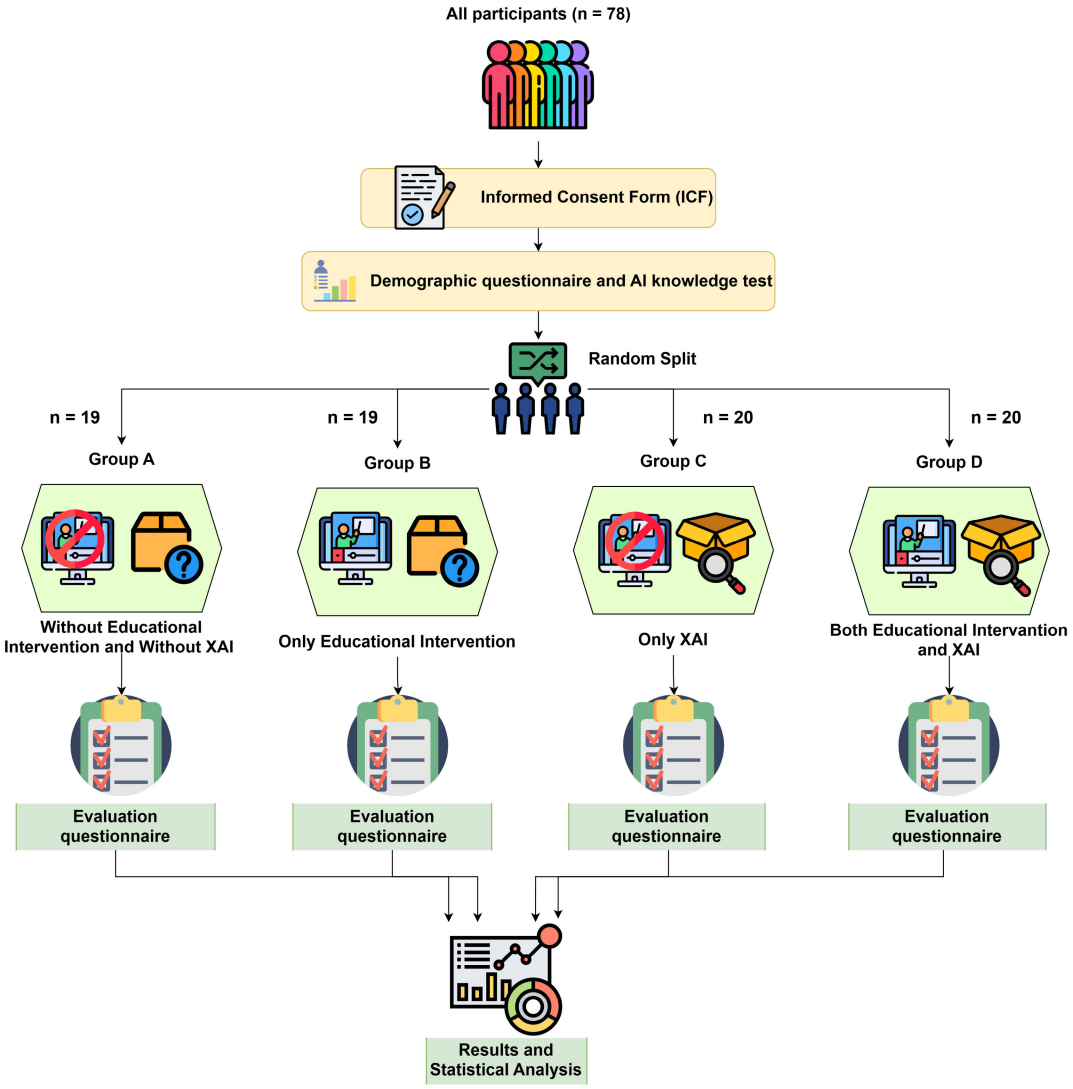


FIGURE 3. Study design.

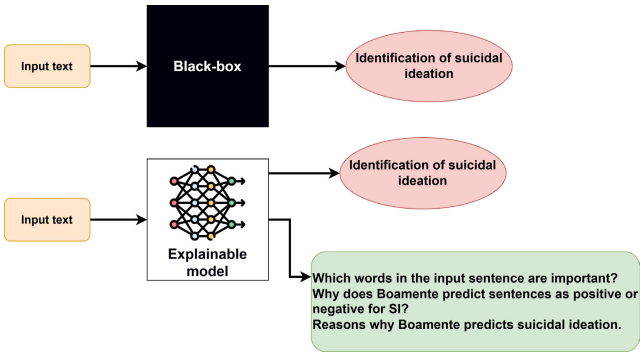


FIGURE 4. *Boamente* without XAI vs. *Boamente* with XAI.

model. At the end of the procedure, participants answered an evaluation questionnaire, which is detailed in the next section. Finally, we analysed the collected data and applied statistical methods.

D. DATA COLLECTION INSTRUMENTS

The first questionnaire applied was the sociodemographic, from which the following were collected: (1) age, (2) type of MHP (psychologist or psychiatrist), (3) clinical experience time, (4) whether he/she had ever treated patients with SI (Yes or No), (5) level of basic computer knowledge, and (6) level of knowledge in AI. Responses to questions (5) and (6) were recorded on a 5-point Likert scale. The second questionnaire, related to the AI knowledge test, assessed the participants' level of AI knowledge. The AI knowledge test was based on questions from the study by Leichtmann et al. [52], with our best translation effort to PT-BR. Each question had four possible answers, of which only one was correct. The test is presented in Table S2 (Supplementary Material) in PT-BR and English.

The evaluation questionnaire included a question on misclassification, a question on trust (Table S3), and ten questions on the quality of explanations (Table S4). However,

the ten questions on explanations were available only to participants in groups C and D. Question 1 had the following statement: "How many sentences did you test? Did you find any sentence classifications inconsistent? Please provide these sentences, along with any others you tested that you believe are relevant to share with us." Question 2 was adapted from the 12° question of the Human-Computer Trust Model (HCTM) [97]. It had the following statement: "Can you trust the information provided by the *Boamente* Artificial Intelligence?". The participants answered the question 2 on a 5-point Likert scale.

The ten questions related to quality of explanations were obtained from the System Causability Scale (SCS) [43]. SCS was designed to swiftly ascertain whether and to what degree an explanation, an explainable user interface, or the explanation process itself is appropriate for the intended application [43]. The final score of the ten questions was calculated from 0 to 100. By using a 5-point Likert scale, the final score is: $SCS_{score} = \frac{\Sigma_i Rating_i}{50}$.

E. ANALYZED FACTORS

We defined the following independent variables (predictor variables) for data analysis:

- Profession: it indicates the type of MHP, namely, psychologist (0) or psychiatrist (1);
- Experience time: it refers to the clinical experience time (in years) of participants;
- Whether the professional treated patients with SI: this variable indicates the value of whether the participant has ever treated a patient with SI (value may be 1 or 0);
- Self-reported basic computing level: it represents a value from 1 to 5, self-reported by participants regarding their level of basic computing knowledge;
- Self-reported AI level: it represents a value from 1 to 5, self-reported by participants regarding their level of AI knowledge;
- AI knowledge test: The score ranges from 0 to 10, based on the participants' performance in an AI knowledge test consisting of 5 questions, each scores 2 points;
- Educational intervention: 1 if the participant had access to the educational video during the experiment, and 0 otherwise;
- Reported misclassification: it records whether the participant reported any sentence misclassification or, in their opinion, the classification confidence probabilities were inadequate. A value of 0 was assigned if no misclassification was reported, and a value of 1 if a misclassification was reported;
- Explanation: 1 if the participant received an explanation during the experiment, and 0 otherwise;
- SCS score: it represents scores ranging from 0 to 100 for the quality of explanations, with a score of zero assigned to groups A and B.

Trust, as measured from question 2, is the dependent variable. Variables were binary, ordinal categorical, and continuous, as follows.

- Binary categorical: profession, whether the professional treated patients with suicidal ideation, with educational intervention, reported misclassification, and with explanation;
- Ordinal categorical: basic computing level (answers from 1 to 5), AI level (answers from 1 to 5), AI knowledge test with 5 questions (each scores 2 points, i.e., from 0 to 10 every two points), and trust (answers from 1 to 5);
- Continuos: clinical experience time and SCS score.

1) STATISTICAL METHODS

Descriptive statistics were conducted to analyse the relationship between groups and trust levels. We conducted assumption tests to determine the applicability of the statistical methods. We analysed the collected data using the tools Jamovi [96], IBM SPSS Statistics [117] and the Python programming language (library statsmodels).

An analysis of covariance (ANCOVA) [30] through the Generalized Linear Model (GLM) [75] was performed to investigate whether there were statistically significant differences in trust levels among groups, controlling for the effect of independent variables. We also performed a GLM analysis to examine the relationship between the AI knowledge test, SCS score, and group on trust, building a prediction model using data exclusively from participants in groups C and D.

GLM is used to find the linear relationship between the response variable (i.e., dependent) and one or more predictor variables (i.e., independent) with different error distributions [79]. ANCOVA analyses aggregated data with the dependent variable and two or more predictor variables (called covariates), where at least one is continuous (quantitative, ordinal) and one is categorical (nominal, non-ordinal) [79], [84]. A GLM analysis was performed in the form of an ANCOVA because it is a method that allows the analysis of quantitative variables, dummy variables corresponding to treatments, design structure, and possible interactions, i.e., it combines regression analysis with analysis of variance (ANOVA) [79]. Furthermore, ANCOVA is an effective method for comparing changes among groups with a random distribution of participants [44].

A post-hoc analysis between the factors With Explanation and With Educational Intervention was conducted following the ANCOVA analysis. Essentially, post-hoc means "after the event", and these analyses were performed to investigate which specific groups differ from each other. While ANCOVA indicates that there is an overall difference, it does not specify where that difference lies [28].

V. RESULTS

A. PARTICIPANT CHARACTERIZATION

Table 1 outlines the sample of 78 participants. Most of the professionals interviewed are predominantly psychologists. This is because participants were recruited from states in

TABLE 1. Sociodemographic data.

Age	**Group A**	**Group B**	**Group C**	**Group D**	**Total**	**%**
18-29	5	5	10	9	**24**	**30.77**
30-39	11	10	7	7	**40**	**51.28**
40-49	3	3	3	4	**13**	**16.67**
50-59	0	1	0	0	**1**	**1.28**
Profession	**Group A**	**Group B**	**Group C**	**Group D**	**Total**	**%**
Psychologist	16	19	18	20	**73**	**93.59**
Psychiatrist	3	0	2	0	**5**	**6.41**
Time of experience	**Group A**	**Group B**	**Group C**	**Group D**	**Total**	**%**
$\geq$ 5 years	12	8	8	10	**38**	**48.72**
$<$ 5 years	7	11	12	10	**40**	**51.28**
Have you ever treated patients with suicidal ideation?	**Group A**	**Group B**	**Group C**	**Group D**	**Total**	**%**
Yes	17	16	17	17	**67**	**85.90**
No	2	3	3	3	**11**	**14.10**
What is your level of knowledge in basic computing?	**Group A**	**Group B**	**Group C**	**Group D**	**Total**	**%**
1 - Very low	0	1	1	0	**2**	**2.56**
2 - Low	3	3	2	1	**9**	**11.54**
3 - Moderate	9	12	10	7	**38**	**48.72**
4 - High	7	3	6	10	**26**	**33.33**
5 - Very high	0	0	1	2	**3**	**3.85**
What is your level of knowledge in artificial intelligence?	**Group A**	**Group B**	**Group C**	**Group D**	**Total**	**%**
1 - Very low	7	8	6	4	**25**	**32.05**
2 - Low	5	6	6	4	**21**	**26.95**
3 - Moderate	4	4	8	10	**26**	**33.33**
4 - High	3	1	0	2	**6**	**7.69**
5 - Very high	0	0	0	0	**0**	**0**
Total					**78**	**100**

the northeast region (Piauí and Ceará), where there is fewer than one psychiatrist per 100,000 inhabitants, and there are more psychologists than other mental health professionals in the country [62]. This fact accounts for the recruitment of more psychologists (n = 73) compared to psychiatrists (n = 5). Approximately half of the participants had over five years of professional experience. Notably, 11 professionals (14.10%) had never treated patients with SI. Additionally, most participants had a moderate or high knowledge of basic computing, but possess low or very low knowledge of AI.

B. DESCRIPTIVE STATISTICS

Figure 5 presents the box plots related to the trust levels of each group or factor. Each box in the box plot represents the distribution of trust data for each group or factor, with the central line indicating the median, the box edges representing the first and third quartiles, and the vertical lines (whiskers) showing the data range, excluding outliers. The black squares within the boxes represent the mean trust level for each group. The numbers below the groups or factors indicate the presence of outliers.

These charts reveal two outliers in groups A and B: participants 18 and 25. There are two approaches to dealing with outliers: (a) removing the outliers and (b) retaining the outliers. The first approach is disadvantageous because it eliminates valuable information. The second approach may result in inaccurate estimates for the model parameters; however, when using robust estimators, GLM models become insensitive to the presence of outliers [68]. Consequently,

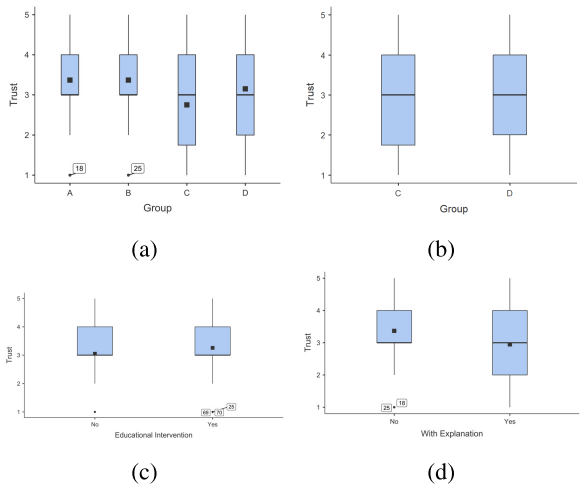


FIGURE 5. Box plots presenting the distribution of trust levels for: (a) four groups of MHPs; (b) groups C and D; (c) groups with and without educational intervention; and (d) groups with and without explanation.

after conducting experiments with and without outliers, we decided to retain the outliers in this study, as they did not significantly influence the GLM results with the robust estimator. Figure 5 shows that there are no outliers in groups C and D.

C. ASSUMPTION TESTING

All assumptions were satisfied in the data analysis for groups A, B, C, and D, as detailed in the Supplementary Material. To determine the applicability of the statistical

methods, several tests were conducted. These included the Kolmogorov-Smirnov and Shapiro-Wilk tests to verify the normality of residuals, the Breusch-Pagan and Levene's tests for homogeneity of residual variances, and the Bartlett test for homoscedasticity. Additionally, collinearity among variables was assessed. All assumptions were met, with no significant deviations observed. Further details and the results of these tests, including Q-Q scatter plot, residual histograms, and Residuals vs. Predictions scatter plots, are presented in Tables S5-S11 and Figures S6-S11 of Supplementary Material.

D. EFFECT ANALYSIS OF FACTORS IN MHP'S TRUST

A GLM analysis was conducted to evaluate the effects of the independent variables on trust, with data distributed into two conditions: With Explanation and With Educational Intervention. The results presented in Table 2 demonstrate that the model was statistically significant, $[F(11) = 3.497, p < 0.001, \eta_p^2 = 0.368]$. The model explained 36.8% of the total variance ($\eta^2 = 0.368$), indicating a large effect size.

In Table 2, the factor With Explanation demonstrated the result of $[F(1) = 6.787, p = 0.011, \eta^2 p = 0.093]$, i.e., it indicates statistically significant differences between groups A and B (without explanation) as compared to groups C and D (with explanation). Furthermore, the variability in MHPs' trust can be attributed to reported misclassification: $[F(1) = 17.557, p < 0.001, \eta^2 p = 0.21]$. A high value suggests greater variability. A high F value indicates that the model explains a significant proportion of the variance. The probability of observing such an extreme result, assuming no real effect, is very small if the p-value is lower than 0.001. The other variables, such as AI knowledge test, educational intervention, self-reported AI level, self-reported basic computing level, clinical experience time, and whether the professional treated patients with SI, did not show a significant effect on professionals' trust, according to the p-values. The interaction between the factors With Explanation and With Educational Intervention was insignificant, indicating that the effect of explainability does not depend on the educational intervention.

The post-hoc test shown in Table 3 revealed an adjusted p-value (p bonferroni = 0.011) and mean difference = 2.815, indicating a significantly higher level of trust of groups A and B, which received no explanation, over groups C and D, which received explanations. There are no significant differences between the groups with educational intervention, and the interaction between conditions is not even significant. The post-hoc test suggests that the group without explanation has a significantly higher mean value than the group with explanation.

The results indicate no relevant differences in the trust level between the groups that received (groups B and D) and those that did not receive (groups A and C) the educational intervention. Participants in groups A and B had similar trust levels. In contrast, the trust level in group D was higher than that of participants in group C, but not significantly. Therefore, the educational intervention had no statistically significant effect on the trust.

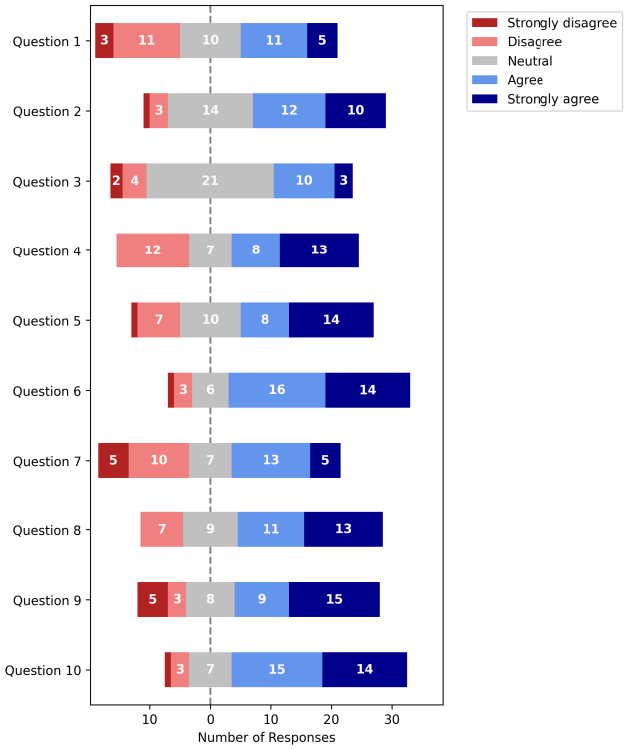


FIGURE 6. Number of responses on the Likert scale for the ten SCS questions.

E. QUALITY OF EXPLANATION AND PRIOR KNOWLEDGE IN AI ON TRUST

The distribution of Likert scale responses from groups C and D (40 responses) regarding the quality of explanations for ten SCS questions (Supplementary Material - Table S4) is shown in Figure 6. Questions 2, 4, 5, 6, and 8 assess whether participants understood the explanations and found them useful for identifying SI in texts. Most participants agreed or strongly agreed with the statements, except for question 4, which had more evenly distributed responses. Question 4 explores whether participants required support to understand the explanations. Notably, for question 5, which concerns understanding the causality of *Boamente* classifying a text as positive or negative for SI, 55% (22 participants) agreed or strongly agreed.

Responses to question 3 indicated a higher concentration of neutral answers, suggesting participants felt indifferent about their ability to adjust the level of detail. For questions 7 and 9, the responses were more evenly distributed, with a slight majority agreeing that the explanations were consistent and complete without needing additional information. Responses to question 10 were similarly distributed, with a slight tendency towards agreement, indicating that some participants

TABLE 2. Results from the ANCOVA method among groups A, B, C, and D.

Variables	Sum of Squares	df	F	p	η^2	$\eta^2 p$	ω^2	$\omega^2 p$	ε^2	$\varepsilon^2 p$
AI knowledge test	0.303	1	0.178	0.675	0.00170	0.003	≤ 0.001			
Self-reported AI Level	0.774	1	0.454	0.503	0.00434	0.007	≤ 0.001			
Self-reported BCL	0.695	1	0.407	0.525	0.00390	0.006	≤ 0.001			
Experience time	0.235	1	0.138	0.712	0.00132	0.002	≤ 0.001			
Treated some patient with SI	0.267	1	0.156	0.694	0.00150	0.002	≤ 0.001			
Reported misclassification	29.953	1	17.557	<.001	0.16806	0.210	0.157	0.175	0.158	0.198
Profession	3.214	1	1.884	0.175	0.01803	0.028	0.008	0.011	0.008	0.013
With Educational Intervention	0.001	1	0.000	0.984	3.88e-6	≤ 0.001				
With Explanation	11.579	1	6.787	0.011	0.06496	0.093	0.055	0.069	0.055	0.080
With Educational Intervention $*$ With Explanation	0.002	1	0.001	0.973	1.14e-5	≤ 0.001				
Note: BCL = basic computing level; df = degrees of freedom; F = F-statistic; p = p-value; η^2 = eta squared; $\eta^2 p$ = partial eta squared; ω^2 = omega squared; $\omega^2 p$ = partial omega squared; ε^2 = epsilon squared; $\varepsilon^2 p$ = partial epsilon squared.										

TABLE 3. Post-hoc comparisons: with explanation and with educational intervention.

Condition	Values		Mean difference	SE	df	t	p bonferroni
With Explanation	No	Yes	2.815	1.080	2.61	66	0.011
With Educational Intervention	No	Yes	0.006	0.319	0.0201	66	0.984
Note: SE = standard error; df = degrees of freedom; t= t-test value; p bonferroni = p-value with Bonferroni correction							

felt they received the explanations in a timely and efficient manner.

To verify whether the quality of explanations (i.e., SCS score) and prior knowledge of AI (i.e., AI knowledge test) had significant effects on trust levels, we performed an ANCOVA with the data collected only from groups C and D. The box plot showed no outliers. Variances were homogeneous, as assessed by the Levene test ($p = 0.759$), and homoscedasticity was met, according to the Bartlett test ($p = 0.40$). Table 4 shows that the ANCOVA test indicated a significant effect of the explanation quality on the trust level [$F(1, 40) = 8.5953, p = 0.006, \eta^2 p = 0.191$]. Prior knowledge of AI measured by the 5-question test and having undergone educational intervention did not present significant differences between the groups.

The effect size η^2 indicates that the SCS score can explain 19.1% of the total variability on trust. The value $\eta^2 p = 0.131$ is similar to η^2, but weighs the significance of the effect. Since $\eta^2 p$ is close to η^2, we can conclude that the effect of SCS score is substantial and statistically significant. The ω^2 indicates that 16.5% of the total variability in the trust level can be attributed exclusively to the covariate SCS score.

The correlation between the SCS score and trust variables, with a Pearson coefficient of 0.457($df = 38, p\text{-}value = 0.003$), indicates a moderate positive relationship. This means that, as the values of the SCS score variable increase, the trust level also tends to increase. The statistical significance ($p < 0.05$) confirms that the relationship is not by chance and is highly likely to be real.

A complementary analysis using GLM was performed to verify whether there is a linear relationship between trust level and the SCS score. According to Table 5, the GLM presented statistical significance ($p = 0.013$) and explained 21.0% ($R^2 = \eta^2 = 0.210$) of the variability of trust variable, with at least one independent variable contributing significantly to the model. Table 5 also shows that the variable SCS score ($p = 0.006$) significantly affects the variable trust. In contrast, the group ($p = 0.867$) does not present a significant effect. Therefore, the data from the analysis confirm that the explanation quality affects trust.

Figure 7 was generated during the correlation analysis between the SCS and trust variables. The graph demonstrates a clear linear relationship between SCS and trust. As SCS increases, the level of trust also rises. Given that SCS measures the quality of explanations produced by the XAI method, the graph suggests that higher quality explanations (indicated by higher SCS values) lead to increased trust in the intelligent system.

VI. DISCUSSION

This exploratory study examined the impact of AI explanations and other factors on trust in the *Boamente* system, an AI-based tool designed to detect SI from texts written in PT-BR. The analysis focused specifically on MHPs, who are the intended end users of the system. While some factors had little to no effect on trust, others significantly influenced the professionals' trust in the system. In this section, we first discuss the findings related to these factors

TABLE 4. Results from the ANCOVA method between groups C and D.

Variables	Sum of Squares	df	Mean Square	F	p	η^2	$\eta^2 p$	ω^2
SCS score	10.8471	1	10.8471	8.5953	0.006	0.191	0.193	0.165
AI knowledge test	0.3229	1	0.3229	0.2559	0.616	0.006	0.007	-0.016
Group	0.0705	1	0.0705	0.0559	0.814	0.001	0.002	-0.021
Note: df = degrees of freedom; F = F-statistic; p = p-value; η^2 = eta squared; $\eta^2 p$ = partial eta squared; ω^2 = omega squared; $\omega^2 p$ = partial omega squared.								

TABLE 5. Results from the linear relationship of SCS score and group on trust level through GLM.

	SS	df	F	p	η^2	$\eta^2 p$	ω^2
Model	12.146	2	4.911	0.013	0.210	0.210	0.164
SCS score	10.546	1	8.528	0.006	0.182	0.187	0.157
Group	0.035	1	0.028	0.867	6.08E-04	0.001	$\leq$ 0.001
Note: SS = Sum of Squares; df = degrees of freedom; F = F-statistic; p = p-value; η^2 = eta squared; $\eta^2 p$ = partial eta squared; ω^2 = omega squared.							

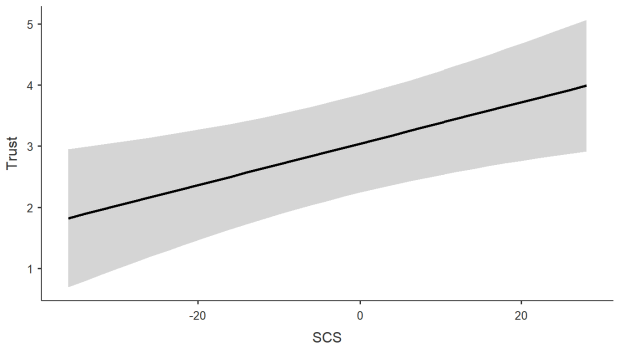


FIGURE 7. Linear relationship: SCS score versus trust.

and their implications. Then, we highlight the strengths of the study while also acknowledging its potential limitations.

A. FACTORS WITHOUT AN EFFECT ON TRUST

In our study, we identified several factors that did not significantly affect trust in the *Boamente* system: educational intervention, AI Knowledge test, profession, clinical experience time, whether the professional treated patients with SI, self-reported basic computing level, and self-reported AI level. Our study examined an educational intervention that had no significant effect. The intervention included a video presentation explaining general concepts of AI, NLP, and XAI, along with an overview of *Boamente*'s functionality in text classification and its use of LIME to explain predictions. This finding aligns with the results reported by Leichtmann et al. [52], where participants received prediction outcomes for the edible mushroom-picking task through a screenshot mock-up of a mobile app. The authors identified this as a methodological limitation that likely contributed to the lack of a effect. Similarly, our study employed a digital video intervention to explain the AI-based system, which also failed to influence participants' trust.

In a study evaluating the effects of two explanation styles on system understanding, persuasive power, and task performance in the context of decision support for diabetes self-management, Waa et al. [110] adopted an educational intervention that included a learning block where multiple stimuli were presented, accompanied by either example-based, rule-based explanations, or no explanation. Different from the results of our study, this educational intervention had a small positive effect on system understanding, but it was not decisive in influencing user trust, as participants exhibited varied behaviours.

We believe that, in our study, an educational intervention solely based on a lengthy video (i.e., 20 minutes) may have constrained the understanding of the system's functionality. Thus, creating shorter videos that incorporate a variety of educational formats could have a greater impact on user trust [48], [52]. Furthermore, we believe that a strongly effect of the educational intervention would be observed if it is based on experimental tasks and simulations, as these methods can effectively explore the limitations of AI. These educational approaches have shown promising results in health training [7], [49].

Our findings indicated that professional experience (i.e., whether the professional treated patients with SI and the years of clinical professional experience), the participant's profession, and self-reported knowledge of basic computer and AI did not impact trust. This finding aligns with the results related by Gaube et al. [37]. The authors found no difference in trust between a group of clinical residents or practising emergency medicine physicians and a group of radiologists in tasks supported by AI for reviewing X-rays; that is, no statistically significant difference between experts and non-experts.

The results also indicated that participants' prior AI knowledge based on an AI test had no largely effect on trust. Similar to our results, Leichtmann et al. [52] suggested that the AI knowledge test could have limited validity due to the study design, as the AI-based application did not allow for exploration. Participants in their study had access only to predetermined images of screenshot, with no opportunity to test the AI or its limitations. However, our results demonstrated that even when participants were given the possibility to explore the proposed tool, factors related to their background did not influence trust. In our study, participants

could explore the interface by classifying different sentences (all groups) and examining explanations (groups C and D). Therefore, MHPs with greater knowledge could have tested *Boamente* by using their expertise in a manner consistent with their usual practices. This outcome from our study, however, differed from the study by Xuan et al. [114], where participants with higher education demonstrated a better understanding of the explanations.

B. FACTORS WITH AN EFFECT ON TRUST

In our study, we found that trust was primarily influenced by the following variables: reported misclassification, with explanation, and SCS score. When participants disagreed with the LLM's classification, probability of the sentence expressing or not SI, or AI explanations, they reduced the level of trust. Some MHPs disagreed with certain probability values of sentences expressing SI or not, perceiving the probability values as too low or inconsistent when compared with the probability values of other sentences. Consequently, when MHPs encountered the vulnerabilities of the LLM and realized that *Boamente* is not infallible, they may have lowered their trust and expectations with the system. In addition, around 55% of participants in the groups with access to explanations (Groups C and D) fully or partially agreed that they understood the reasons why the *Boamente* system identifies SI in texts. Therefore, the findings suggest a loss of expectations from the LLM after participants understand how it internally classifies sentences and is susceptible to incorrect predictions.

LIME-based explanations had a moderate effect size on the trust of participants, while the effect size of report misclassification was higher. Initially, we believed that LIME-based explanations would have a largely effect on achieving higher trust in the participants in groups C and D compared to those in groups A and B. However, the explanations decreased MHPs' trust: participants in groups A and B had higher trust than those in groups C and D. That is, participants without access to the explanations exhibited a slightly higher level of trust compared to those with access to the explanations. This finding is consistent with the primary studies conducted in [11], [52], [110], and [114].

This study focused on the trust assessment of a SI detection tool by MHPs, who are inherently rigorous users compared to other types of AI-based systems. The participants approached the *Boamente* with the intention of using it to care for their patients with SI. This professional context heightened their scrutiny and expectations from the system, as their primary concern was ensuring patient safety and well-being. Then, these professionals perceived the use of the *Boamente* as a critical decision-making process, essentially placing their patients' lives in the hands of the system. This perspective impacts their interaction with the tool, requiring it to be highly reliable and accurate. Additionally, when using a simple classification interface without explanations of predictions, participants in groups A and B are likely to view the system as functioning satisfactorily because they assumed it met the objective of correctly classifying most of the sentences tested. MHPs using a straightforward interface (groups A and B) tend to be less concerned with the reasoning or internal mode of operation and more focused on the system's accuracy [116]. Even when AI performs poorly, users might misattribute the blame to themselves [24], [114].

The *Boamente* is a tool designed to identify SI and serve as an alert system for potentially severe outcomes. Alerts necessitate careful consideration, as they serve to notify the professional that a patient requires timely and appropriate intervention. We highlight that professionals might adopt a mindset where, even if the system occasionally errs, it is preferable to always intervene more closely with the patient, then providing increased attention (e.g., through conversations, therapy sessions, or even app messages). Trust in the system can be established with an understanding of its limitations and the acceptance that a few errors might occur. Therefore, professionals should consider the tool as a supplementary alert mechanism (i.e., not a ''primary solution'') that prompts them to take necessary actions, thus improving the patient's safety and possibly preventing adverse outcomes.

When considering only the data from groups C and D, which had access to the explanations, the SCS scores had a high effect size on trust. Therefore, we conclude that the quality attributed to the explanations by the participants significantly impacts their trust. Specifically, higher-quality explanations reported by MHPs corresponded to higher levels of trust. The SCS score accounted for a substantial portion of the variability in trust (approximately 19%), indicating a positive effect. Therefore, if LIME-based explanations provided to groups C and D were perceived as overly complex, confusing, or uninformative, they may have diminished participants' trust.

The literature generally indicates that methods provided by the XAI field are the recommended solutions to mitigate trust issues and improve the acceptance of AI in healthcare applications. Specifically, XAI has the potential to enhance the trust of healthcare professionals in AI-based systems [2], [3], [6], [23], [85], [94]. However, the effectiveness of explanations in building trust varies. Some studies show an increase in trust levels [17], [72], [87], [94], [116], while others present contradictory results, indicating that providing explanations may not increase satisfaction or might even decrease users' trust in a system [27], [37], [45], [51], [94], [116]. Zhang et al. [116] concluded that local explanations of the logic of an AI prediction enhanced users' understanding, but did not necessarily increase trust. Rosenbacke et al. [94] pointed to an increase in clinicians' trust in XAI vs. standard AI, as well no significant effect of XAI on trust, indicating that the presence of explanations does not automatically improve trust.

That variation in the effect of XAI methods on trust can be attributed to the styles and, most importantly, the quality of the explanations. Xuan et al. [114] found that feature importance and decision boundary visualization were styles

of explanations the most understandable, but users often failed to recognize their limitations, leading to overtrust. Correctly interpreting explanations was linked to high trust, while failing to assess their limits resulted in overtrust. The study by Xuan et al. [114] also revealed that explanations deemed easy to understand could be more misleading. In line with the findings by Rosenbacke et al. [94] about XAI vs. trust, complex or contradictory explanations can undermine trust in AI. They underscore the nuanced role of explanation quality and suggest that trust can be modulated through the careful design of XAI systems.

Based on the above considerations, we suppose that the reason for a lower trust level in groups C and D compared to groups A and B may stem from the use of LIME-based explanations for AI predictions. These explanations likely fostered more adequate (or optimal) trust among MHPs in groups C and D. The trust of humans is optimal when there is some level of balanced skepticism regarding the AI's decisions to avoid errors that might arise from excessive under- or overtrust [11], [94]. When trust is at its maximum, the user accepts or believes all recommendations and results generated by the AI system with low or no scepticism [11]. Hence, maintaining a balance between trust and scepticism is crucial to prevent AI models from diminishing the quality of patient care [25]. Excessive trust — such as trusting AI-generated advice without a healthy level of critical thinking — can harm clinical accuracy if the advice is incorrect. Similarly, even correct advice may be disregarded due to user distrust, leading to the same negative outcome [94].

We conclude that the quality of explanations provided by the *Boamente* system is critical in fostering an appropriate level of trust among MHPs. Simply providing explanations does not automatically increase trust, as shown by the lower trust levels reported in groups that received LIME-based explanations (groups C and D) compared to those that did not (groups A and B). This highlights the importance of high-quality, clear explanations to avoid both excessive trust and unwarranted distrust. Moreover, striking the right balance between trust and scepticism is essential for clinical accuracy in identifying SI in texts. This balance ensures that patients at risk of SI who are not identified do not go without adequate care, while also preventing those correctly identified from being unfairly overlooked.

C. STRENGTHS AND LIMITATIONS

This study addressed the importance of building trust between humans and AI, particularly in the context of mental health. We emphasize the rigour of its methodology, which included the recruitment of 78 psychologists and psychiatrists, ensuring each participant was randomly assigned to a specific group. The experiment was conducted with the system's end-users, i.e., MHPs who are potential future users of the system. This ensures that the evaluation was carried out by experts in mental health, thereby enhancing the relevance and applicability of the findings. An additional strength of this study is the utilization of a system in production, allowing MHPs to interact with it by simulating patient inputs. This approach is superior to mock-ups or static interfaces, as it provides users with the opportunity to experiment with different sentences. So, participants had access to functional interfaces that enabled them to type and classify sentences.

This study has limitations to be acknowledged. First, our study focused on a video-based explanation and two XAI methods: the LIME technique, which provides visual explanations based on feature importance, and an example-based approach that uses positive and negative sentence examples to provide explanation. Rule-based explanations and other XAI methods were not evaluated, which could provide new insights into the effects of explanations on understanding and trust in AI. Exploring these approaches could help identify which explanation styles are most effective for fostering appropriate trust and comprehension for MHPs. Second, the recruitment was confined to only two Brazilian states. This geographical constraint implies that the findings may not be fully representative of the broader Brazilian population. Finally, a limitation of this study is the lack of recruitment of MHPs with backgrounds other than psychologists and psychiatrists who also work with mental health and suicide prevention, such as nurses and social workers. Including these professionals could provide a more comprehensive and diverse perspective, enriching the analysis of trust in AI explanations and their applicability in different contexts of mental health practice.

VII. CONCLUSION

This study aimed to examine how MHPs' trust is affected when using the *Boamente* system. It explored mainly AI explanations (educational intervention and XAI methods) and other factors related to professional background, knowledge of AI and computing, and reported system misclassification. Our study evaluated the impact of factors on the trust of MHPs across four groups of participants. Specifically, we employed the LIME method to elucidate the predictions of the BERTimbau-Large model. We observed that participants in groups without access to explanations exhibited slightly greater trust than MHPs with access to explanations. Furthermore, disagreement with AI classifications and perceptions of system vulnerabilities also affected trust. The results indicated that factors related to the professional background, knowledge of AI and computing, and educational intervention did not affect the trust. Explaining predictions can lead to more realistic and informed trust in AI systems.

As part of future work, we plan to evaluate the impact of different XAI methods on the understanding and trust of MHPs. We also plan to fine-tune generative LLMs to not only predict SI but also provide textual explanations for these predictions. These evaluations will enable us to analyse how different explanations impact MHPs' trust in the system. Unlike the LIME-based explanations examined in this study, textual explanations for SI predictions in non-clinical texts represent a distinct approach and are possibly more easily

understood by MHPs. Our aim will be to assess whether these explanations can enhance MHPs' trust and improve their understanding of AI-generated predictions.

DATA AVAILABILITY

The *Boamente* dataset is available at https://doi.org/10.5281/zenodo.10070747. The code used is available at https://github.com/ adonias-caetano/ufdpar-human-factors-xai-trust.git

REFERENCES

[1] A. Abid, A. Abdalla, A. Abid, D. Khan, A. Alfozan, and J. Zou, "Gradio: Hassle-free sharing and testing of ML models in the wild," 2019, *arXiv:1906.02569*.

[2] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[3] A. Adadi and M. Berrada, "Explainable AI for healthcare: From black box to interpretable models," in *Embedded Systems and Artificial Intelligence*, V. Bhateja, S. C. Satapathy, and H. Satori, Eds., Singapore: Springer, 2020, pp. 327–337.

[4] A. Adjekum, A. Blasimme, and E. Vayena, "Elements of trust in digital health systems: Scoping review," *J. Med. Internet Res.*, vol. 20, no. 12, Dec. 2018, Art. no. e11254.

[5] A. E. Aladağ, S. Muderrisoglu, N. B. Akbas, O. Zahmacioglu, and H. O. Bingol, "Detecting suicidal ideation on forums: Proof-of-concept study," *J. Med. Internet Res.*, vol. 20, no. 6, Jun. 2018, Art. no. e215.

[6] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101805.

[7] A. A. Vanderbilt, N. J. Pastis, J. Mayglothling, and D. Franzen, "A review of the literature: Direct and video laryngoscopy with simulation as educational intervention," *Adv. Med. Educ. Pract.*, vol. 5, pp. 15–23, Jan. 2014.

[8] R. Alonzo, J. Hussain, S. Stranges, and K. K. Anderson, "Interplay between social media use, sleep quality, and mental health in youth: A systematic review," *Sleep Med. Rev.*, vol. 56, Apr. 2021, Art. no. 101414.

[9] American Psychiatric Association, *DSM-5: Diagnostic and Statistical Manual of Mental Disorders*, A. Editora, Ed., Washington, DC, USA: American Psychiatric Association, 2014.

[10] A. Arowosegbe and T. Oyelade, "Application of natural language processing (NLP) in detecting and preventing suicide ideation: A systematic review," *Int. J. Environ. Res. Public Health*, vol. 20, no. 2, p. 1514, Jan. 2023.

[11] O. Asan, A. E. Bayrak, and A. Choudhury, "Artificial intelligence and human trust in healthcare: Focus on clinicians," *J. Med. Internet Res.*, vol. 22, no. 6, Jun. 2020, Art. no. e15154.

[12] L. Balcombe and D. De Leo, "Digital mental health challenges and the horizon ahead for solutions," *JMIR Mental Health*, vol. 8, no. 3, Mar. 2021, Art. no. e26811.

[13] S. Booniam, T. Wongpakaran, P. Lerttrakarnnon, S. Jiraniramai, P. Kuntawong, and N. Wongpakaran, "Predictors of passive and active suicidal ideation and suicide attempt among older people: A study in tertiary care settings in Thailand," *Neuropsychiatric Disease Treatment*, vol. 16, pp. 3135–3144, Dec. 2020.

[14] A. Bourla, F. Ferreri, L. Ogorzelec, C.-S. Peretti, C. Guinchard, and S. Mouchabac, "Psychiatrists' attitudes toward disruptive new technologies: Mixed-methods study," *JMIR Mental Health*, vol. 5, no. 4, Dec. 2018, Art. no. e10240.

[15] A. Bourla, S. Mouchabac, L. Ogorzelec, C. Guinchard, and F. Ferreri, "Are student nurses ready for new technologies in mental health? Mixed-methods study," *Nurse Educ. Today*, vol. 84, Jan. 2020, Art. no. 104240.

[16] J. M. Braciszewski, "Digital technology for suicide prevention," *Adv. Psychiatry Behav. Health*, vol. 1, no. 1, pp. 53–65, Sep. 2021.

[17] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman, "Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems," in *Proc. 25th Int. Conf. Intell. User Interfaces*. New York, NY, USA: Association for Computing Machinery, Mar. 2020, pp. 454–464.

[18] L. Cao, H. Zhang, and L. Feng, "Building and using personal knowledge graph to improve suicidal ideation detection on social media," *IEEE Trans. Multimedia*, vol. 24, pp. 87–102, 2022.

[19] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, Jul. 2019.

[20] G. Castillo-Sánchez, G. Marques, E. Dorronzoro, O. Rivera-Romero, M. Franco-Martín, and I. De La Torre-Díez, "Suicide risk assessment using machine learning and social networks: A scoping review," *J. Med. Syst.*, vol. 44, no. 12, p. 205, Nov. 2020.

[21] *Conselho Federal De Psicologia CFG*, Federal Council Psychol., Porto Alegre, Brazil, Sep. 2023.

[22] *Busca Por Médicos, Conselho Federal De Medicina CFM*, Federal Council Med., Porto Alegre, Brazil, Sep. 2023.

[23] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (XAI)," *IEEE Access*, vol. 11, pp. 78994–79015, 2023.

[24] L. Chong, G. Zhang, K. Goucher-Lambert, K. Kotovsky, and J. Cagan, "Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice," *Comput. Hum. Behav.*, vol. 127, Sep. 2021, Art. no. 107018.

[25] A. Choudhury and Z. Chaudhry, "Large language models and user trust: Consequence of self-referential learning loop and the deskilling of health care professionals," *J. Med. Internet Res.*, vol. 26, Apr. 2024, Art. no. e56764.

[26] H. Cortés, M. Rojas-Márquez, M. L. Del Prado-Audelo, O. D. Reyes-Hernández, M. González-Del Carmen, and G. Leyva-Gómez, "Alterations in mental health and quality of life in patients with skin disorders: A narrative review," *Int. J. Dermatol.*, vol. 61, no. 7, pp. 783–791, Jul. 2022.

[27] H. Cramer, V. Evers, S. Ramlal, M. van Someren, L. Rutledge, N. Stash, L. Aroyo, and B. Wielinga, "The effects of transparency on trust in and acceptance of a content-based art recommender," *User Model. User-Adapted Interact.*, vol. 18, no. 5, pp. 455–496, Nov. 2008.

[28] D. Curran-Everett and H. Milgrom, "Post-hoc data analysis: Benefits and limitations," *Current Opinion Allergy Clin. Immunol.*, vol. 13, no. 3, pp. 223–224, 2013.

[29] D. De Leo, "Late-life suicide in an aging world," *Nature Aging*, vol. 2, no. 1, pp. 7–12, Jan. 2022.

[30] M. B. de Melo, D. Daldegan-Bueno, M. G. M. Oliveira, and A. L. de Souza, "Beyond ANOVA and MANOVA for repeated measures: Advantages of generalized estimated equations and generalized linear mixed models and its use in neuroscience research," *Eur. J. Neurosci.*, vol. 56, no. 12, pp. 6089–6098, Dec. 2022.

[31] A. C. de Oliveira, E. J. S. Diniz, S. Teixeira, and A. S. Teles, "How can machine learning identify suicidal ideation from user's texts? Towards the explanation of the boamente system," *Proc. Comput. Sci.*, vol. 206, pp. 141–150, Jan. 2022.

[32] E. J. S. Diniz, J. E. Fontenele, A. C. de Oliveira, V. H. Bastos, S. Teixeira, R. L. Rabêlo, D. B. Calçada, R. M. dos Santos, A. K. de Oliveira, and A. S. Teles, "Boamente: A natural language processing-based digital phenotyping tool for smart monitoring of suicidal ideation," *Healthcare*, vol. 10, no. 4, p. 698, Apr. 2022.

[33] J. A. Duell, "A comparative approach to explainable artificial intelligence methods in application to high-dimensional electronic health records: Examining the usability of XAI," 2021, *arXiv:2103.04951*.

[34] *Eli5.sklearn.explain_prediction (DOCs API)*, ELI5, 2016. Accessed: May 10, 2024. [Online]. Available: https://eli5.readthedocs.io/en/latest/autodocs/sklearn.html

[35] R. P. Evans, L. D. Bryant, G. Russell, and K. Absolom, "Trust and acceptability of data-driven clinical recommendations in everyday practice: A scoping review," *Int. J. Med. Informat.*, vol. 183, Mar. 2024, Art. no. 105342.

[36] A. Fan, Y. Jernite, E. Perez, D. Grangier, J. Weston, and M. Auli, "ELI5: Long form question answering," *CoRR*, vol. abs/1907.09190, pp. 1–17, Jan. 2019.

[37] S. Gaube, H. Suresh, M. Raue, E. Lermer, T. K. Koch, M. F. C. Hudecek, A. D. Ackery, S. C. Grover, J. F. Coughlin, D. Frey, F. C. Kitamura, M. Ghassemi, and E. Colak, "Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays," *Sci. Rep.*, vol. 13, no. 1, p. 1383, Jan. 2023.

[38] M. G. Gelder, J. J. Lopez-Ibor, and N. C. Andreasen, *New Oxford Textbook of Psychiatry*, vol. 2. London, U.K.: Oxford Univ. Press, 2003.

[39] F. Gille, A. Jobin, and M. Ienca, "What we talk about when we talk about trust: Theory of trust for AI in healthcare," *Intell.-Based Med.*, vols. 1–2, Nov. 2020, Art. no. 100001.

[40] S. D. Gosling and W. Mason, "Internet research in psychology," *Annu. Rev. Psychol.*, vol. 66, no. 1, pp. 877–902, Jan. 2015.

[41] C. M. Greco, A. Simeri, A. Tagarelli, and E. Zumpano, "Transformer-based language models for mental health issues: A survey," *Pattern Recognit. Lett.*, vol. 167, pp. 204–211, Mar. 2023.

[42] B. Harmer, S. Lee, T. V. H. Duong, and A. Saadabadi, Suicidal Ideation. Treasure Island, FL, USA: StatPearls Publishing, 2021. [Online]. Available: http://europepmc.org/books/NBK565877

[43] A. Holzinger, A. Carrington, and H. Müller, "Measuring the quality of explanations: The system causability scale (SCS): Comparing human and machine explanations," *Kunstliche Intelligenz*, vol. 34, no. 2, pp. 193–198, Jan. 2020.

[44] J. Jamieson, "Analysis of covariance (ANCOVA) with difference scores," *Int. J. Psychophysiology*, vol. 52, no. 3, pp. 277–283, May 2004.

[45] R. F. Kizilcec, "How much information? Effects of transparency on trust in an algorithmic interface," in *Proc. CHI Conf. Human Factors Comput. Syst.* New York, NY, USA: Association for Computing Machinery, May 2016, pp. 2390–2395.

[46] J. Kraus, D. Scholz, D. Stiegemeier, and M. Baumann, "The more you know: Trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 62, no. 5, pp. 718–736, Aug. 2020.

[47] A. Kumar, A. Guleria, Sunita, and G. Singh, "Overview of AI based e-healthcare system," *EPRA Int. J. Multidisciplinary Res.*, vol. 9, no. 6, pp. 182–186, Jun. 2023. [Online]. Available: https://eprajournals.com/IJMR/article/10815

[48] M. Körber, E. Baseler, and K. Bengler, "Introduction matters: Manipulating trust in automation and reliance in automated driving," *Appl. Ergonom.*, vol. 66, pp. 18–31, Jan. 2018.

[49] M. Lavelle, C. Attoe, C. Tritschler, and S. Cross, "Managing medical emergencies in mental health settings using an interprofessional in-situ simulation training programme: A mixed methods evaluation study," *Nurse Educ. Today*, vol. 59, pp. 103–109, Dec. 2017.

[50] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 46, no. 1, pp. 50–80, 2004.

[51] B. Leichtmann, A. Hinterreiter, C. Humer, M. Streit, and M. Mara, "Explainable artificial intelligence improves human decision-making: Results from a mushroom picking experiment at a public art festival," *Int. J. Hum.-Comput. Interact.*, vol. 40, no. 17, pp. 4787–4804, Sep. 2024.

[52] B. Leichtmann, C. Humer, A. Hinterreiter, M. Streit, and M. Mara, "Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task," *Comput. Hum. Behav.*, vol. 139, Feb. 2023, Art. no. 107539.

[53] D. Lekkas, R. J. Klein, and N. C. Jacobson, "Predicting acute suicidal ideation on Instagram using ensemble machine learning models," *Internet Intervent.*, vol. 25, Sep. 2021, Art. no. 100424.

[54] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020.

[55] D. Long and B. Magerko, "What is AI literacy? Competencies and design considerations," in *Proc. CHI Conf. Hum. Factors Comput. Syst.* New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–16.

[56] M. Lotito and E. Cook, "A review of suicide risk assessment instruments and approaches," *Mental Health Clinician*, vol. 5, no. 5, pp. 216–223, Sep. 2015.

[57] O. Loyola-González, "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view," *IEEE Access*, vol. 7, pp. 154096–154113, 2019.

[58] J. Luby and S. Kertz, "Increasing suicide rates in early adolescent girls in the United States and the equalization of sex disparity in suicide: The need to investigate the role of social media," *J. Amer. Med. Assoc. Netw. Open*, vol. 2, no. 5, May 2019, Art. no. e193916.

[59] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *CoRR*, vol. abs/1705.07874, pp. 1–10, Jan. 2017.

[60] A. Madsen, S. Reddy, and S. Chandar, "Post-hoc interpretability for neural NLP: A survey," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–42, Dec. 2022.

[61] A. Malhotra and R. Jindal, "XAI transformer based approach for interpreting depressed and suicidal user behavior on online social networks," *Cognit. Syst. Res.*, vol. 84, Mar. 2024, Art. no. 101186.

[62] M. D. Mateus, J. J. Mari, P. G. Delgado, N. Almeida-Filho, T. Barrett, J. Gerolin, S. Goihman, D. Razzouk, J. Rodriguez, R. Weber, S. B. Andreoli, and S. Saxena, "The mental health system in Brazil: Policies and future challenges," *Int. J. Mental Health Syst.*, vol. 2, no. 1, p. 12, Sep. 2008.

[63] J. A. McDermid, Y. Jia, Z. Porter, and I. Habli, "Artificial intelligence explainability: The technical and ethical dimensions," *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 379, no. 2207, Oct. 2021, Art. no. 20200363.

[64] J. Melcher, R. Hays, and J. Torous, "Digital phenotyping for mental health of college students: A clinical review," *Evidence Based Mental Health*, vol. 23, no. 4, pp. 161–166, Nov. 2020.

[65] J. Mendes, F. Silva, A. Cardoso, I. Moura, L. Coutinho, D. Viana, M. Endler, and A. S. Teles, "OpenDPMH: A framework for developing mobile sensing applications of digital phenotyping," in *Proc. IEEE 36th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2023, pp. 198–203.

[66] J. P. M. Mendes, I. R. Moura, P. Van De Ven, D. Viana, F. J. S. Silva, L. R. Coutinho, S. Teixeira, J. J. P. C. Rodrigues, and A. S. Teles, "Sensing apps and public data sets for digital phenotyping of mental health: Systematic review," *J. Med. Internet Res.*, vol. 24, no. 2, Feb. 2022, Art. no. e28735.

[67] P. Mishra, *Practical Explainable AI Using Python: Artificial Intelligence Model Explanations Using Python-based Libraries, Extensions, and Frameworks*. New York, NY, USA: Apress, 2022.

[68] H. R. Moheghi, R. Noorossana, and O. Ahmadi, "GLM profile monitoring using robust estimators," *Qual. Rel. Eng. Int.*, vol. 37, no. 2, pp. 664–680, Mar. 2021.

[69] D. C. Mohr, K. Shilton, and M. Hotopf, "Digital phenotyping, behavioral sensing, or personal sensing: Names and transparency in the digital age," *NPJ Digit. Med.*, vol. 3, no. 1, p. 45, Mar. 2020.

[70] C. Molnar, *Interpretable Machine Learning*. Victoria, BC, Canada: Leanpub, 2020.

[71] I. Moura, A. Teles, F. Silva, D. Viana, L. Coutinho, F. Barros, and M. Endler, "Mental health ubiquitous monitoring supported by social situation awareness: A systematic review," *J. Biomed. Informat.*, vol. 107, Jul. 2020, Art. no. 103454.

[72] J. Muramatsu and W. Pratt, "Transparent queries: Investigation users' mental models of search engines," in *Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.* New York, NY, USA: Association for Computing Machinery, Sep. 2001, pp. 217–224.

[73] A. Naghavi, T. Teismann, Z. Asgari, M. R. Mohebbian, M. Mansourian, and M. Á. Mañanas, "Accurate diagnosis of suicide ideation/behavior using robust ensemble machine learning: A university student population in the middle east and North Africa (MENA) region," *Diagnostics*, vol. 10, no. 11, p. 956, Nov. 2020.

[74] J. Neeleman, R. de Graaf, and W. Vollebergh, "The suicidal process; prospective comparison between early and later stages," *J. Affect. Disorders*, vol. 82, no. 1, pp. 43–52, Oct. 2004.

[75] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *Roy. Stat. Soc. J. A, Gen.*, vol. 135, no. 3, pp. 370–384, Dec. 2018.

[76] D. T. K. Ng, J. K. L. Leung, K. W. S. Chu, and M. S. Qiao, "AI literacy: Definition, teaching, evaluation and ethical issues," *Proc. Assoc. Inf. Sci. Technol.*, vol. 58, no. 1, pp. 504–509, Oct. 2021.

[77] S. D. Nielsen, R. H. B. Christensen, T. Madsen, K. Karstoft, L. K. H. Clemmensen, and M. E. Benros, "Prediction models of suicide and non-fatal suicide attempt after discharge from a psychiatric inpatient stay: A machine learning approach on nationwide Danish registers," *Acta Psychiatrica Scandinavica*, vol. 148, no. 6, pp. 525–537, Nov. 2023.

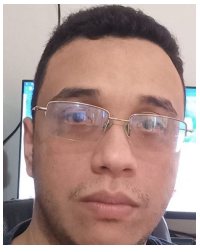
[78] N. Nordin, Z. Zainol, M. H. M. Noor, and L. F. Chan, "An explainable predictive model for suicide attempt risk using an ensemble learning and Shapley additive explanations (SHAP) approach," *Asian J. Psychiatry*, vol. 79, Jan. 2023, Art. no. 103316.

[79] U. Olsson, *Generalized Linear Models: An Applied Approach*. La Vergne, TN, USA: Lightning Source, 2002.

[80] *Mental Health Action Plan 2013–2020*, World Health Org., Geneva, Switzerland, Jan. 2013.

[81] *World Mental Health Report: Transforming Mental Health for All, 2022*, World Health Org., Geneva, Switzerland, 2022.

[82] U. Pagallo, S. O'Sullivan, N. Nevejans, A. Holzinger, M. Friebe, F. Jeanquartier, C. Jean-Quartier, and A. Miernik, "The underuse of AI in the health sector: Opportunity costs, success stories, risks and recommendations," *Health Technol.*, vol. 14, no. 1, pp. 1–14, Jan. 2024.
[83] S. Pengpid and K. Peltzer, "The prevalence and correlates of suicidal ideation, plans and suicide attempts among 15- to 69-year-old persons in Eswatini," *Behav. Sci.*, vol. 10, no. 11, p. 172, Nov. 2020.
[84] D. Philippas, *Analysis of Covariance (ANCOVA)*. Cham, Switzerland: Springer, 2023, pp. 179–183.
[85] D. Pradhan, A. Varshney, and S. Singh, "A critical review on challenges and limitations in artificial intelligence-based e-health applications," Juniper, Tech. Rep. 001-005, May 2023, vol. 5. [Online]. Available: 10.19080/ETOAJ.2023.05.555654
[86] M. Prince, V. Patel, S. Saxena, M. Maj, J. Maselko, M. R. Phillips, and A. Rahman, "No health without mental health," *Lancet*, vol. 370, no. 9590, pp. 859–877, 2007.
[87] P. Pu and L. Chen, "Trust building with explanation interfaces," in *Proc. 11th Int. Conf. Intell. User Interfaces*. New York, NY, USA: Association for Computing Machinery, Jan. 2006, pp. 93–100.
[88] S. T. Rabani, Q. R. Khan, and A. M. U. D. Khanday, "Detection of suicidal ideation on Twitter using machine learning & ensemble approaches," *Baghdad Sci. J.*, vol. 17, no. 4, p. 1328, Dec. 2020.
[89] A. Rawal, J. McCoy, D. B. Rawat, B. M. Sadler, and R. S. Amant, "Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives," *IEEE Trans. Artif. Intell.*, vol. 3, no. 6, pp. 852–866, Dec. 2022.
[90] J. Renaud, S. L. MacNeil, L. Vijayakumar, M. Spodenkiewicz, S. Daniels, D. A. Brent, and G. Turecki, "Suicidal ideation and behavior in youth in low- and middle-income countries: A brief review of risk factors and implications for prevention," *Frontiers Psychiatry*, vol. 13, Dec. 2022, Art. no. 1044354.
[91] C. F. Reynolds, D. V. Jeste, P. S. Sachdev, and D. G. Blazer, "Mental health care for older adults: Recent advances and new directions in clinical practice and research," *World Psychiatry*, vol. 21, no. 3, pp. 336–363, Oct. 2022.
[92] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1135–1144.
[93] J. Robinson, G. Cox, E. Bailey, S. Hetrick, M. Rodrigues, S. Fisher, and H. Herrman, "Social media and suicide prevention: A systematic review," *Early Intervent Psychiatry*, vol. 10, no. 2, pp. 103–121, Apr. 2016.
[94] R. Rosenbacke, Å. Melhus, M. McKee, and D. Stuckler, "How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: Systematic review," *JMIR AI*, vol. 3, Oct. 2024, Art. no. e53207.
[95] B. N. Rudd and R. S. Beidas, "Digital mental health: The answer to the global mental health crisis?" *JMIR Ment Healt*, vol. 7, no. 6, p. 2, Jun. 2020, Art. no. e18472.
[96] M. Şahin and E. Aybek, "Jamovi: An easy to use statistical software for the social scientists," *Int. J. Assessment Tools Educ.*, vol. 6, no. 4, pp. 670–692, Jan. 2020.
[97] S. Gulati, S. Sousa, and D. Lamas, "Design, development and evaluation of a human-computer trust scale," *Behav. Inf. Technol.*, vol. 38, no. 10, pp. 1004–1015, Oct. 2019.
[98] D. Sikander, M. Arvaneh, F. Amico, G. Healy, T. Ward, D. Kearney, E. Mohedano, J. Fagan, J. Yek, A. F. Smeaton, and J. Brophy, "Predicting risk of suicide using resting state heart rate," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2016, pp. 1–4.
[99] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: Pretrained BERT models for Brazilian Portuguese," in *Intelligent Systems*, R. Cerri and R. C. Prati, Eds., Cham, Switzerland: Springer, 2020, pp. 403–417.
[100] F. C. Souza, R. F. Nogueira, and R. A. Lotufo, "BERT models for Brazilian Portuguese: Pretraining, evaluation and tokenization analysis," *Appl. Soft Comput.*, vol. 149, Dec. 2023, Art. no. 110901.
[101] D. J. Stein, A. C. Palk, and K. S. Kendler, "What is a mental disorder? An exemplar-focused approach," *Psychol. Med.*, vol. 51, no. 6, pp. 894–901, Apr. 2021.
[102] H. Tang, A. Miri Rekavandi, D. Rooprai, G. Dwivedi, F. M. Sanfilippo, F. Boussaid, and M. Bennamoun, "Analysis and evaluation of explainable artificial intelligence on suicide risk assessment," *Sci. Rep.*, vol. 14, no. 1, p. 6163, Mar. 2024.
[103] M. Tanguay-Sela et al., "Evaluating the perceived utility of an artificial intelligence-powered clinical decision support system for depression treatment using a simulation center," *Psychiatry Res.*, vol. 308, Feb. 2022, Art. no. 114336.
[104] T. M. Brill, L. Munoz, and R. J. Miller, "Siri, alexa, and other digital assistants: A study of customer satisfaction with artificial intelligence applications," *J. Marketing Manage.*, vol. 35, nos. 15–16, pp. 1401–1436, Oct. 2019.
[105] J. Torous, M. V. Kiang, J. Lorme, and J.-P. Onnela, "New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research," *JMIR Mental Health*, vol. 3, no. 2, p. e16, May 2016.
[106] H. Tremain, C. McEnery, K. Fletcher, and G. Murray, "The therapeutic alliance in digital mental health interventions for serious mental illnesses: Narrative review," *JMIR Mental Health*, vol. 7, no. 8, Aug. 2020, Art. no. e17204.
[107] V. Tucci, J. Saary, and T. E. Doyle, "Factors influencing trust in medical artificial intelligence for healthcare professionals: A narrative review," *J. Med. Artif. Intell.*, vol. 5, p. 4, Mar. 2022.
[108] G. Turecki and D. A. Brent, "Suicide and suicidal behaviour," *Lancet*, vol. 387, no. 10024, pp. 1227–1239, Sep. 2015.
[109] B. H. M. van der Velden, "Explainable AI: Current status and future potential," *Eur. Radiol.*, vol. 34, no. 2, pp. 1187–1189, Aug. 2023.
[110] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, "Evaluating XAI: A comparison of rule-based and example-based explanations," *Artif. Intell.*, vol. 291, Feb. 2021, Art. no. 103404.
[111] V. Venkatesh, J. Y. L. Thong, F. K. Y. Chan, P. J.-H. Hu, and S. A. Brown, "Extending the two-stage information systems continuance model: Incorporating UTAUT predictors and the role of context," *Inf. Syst. J.*, vol. 21, no. 6, pp. 527–555, Nov. 2011.
[112] A. Wongkoblap, M. A. Vadillo, and V. Curcin, "Researching mental health disorders in the era of social media: Systematic review," *J. Med. Internet Res.*, vol. 19, no. 6, p. e228, Jun. 2017.
[113] World Health Organization. (2021). *Suicide Worldwide in 2019: Global Health Estimates*. [Online]. Available: https://www.who.int/teams/mental-health-and-substance-use/data-research/suicide-data
[114] Y. Xuan, E. Small, K. Sokol, D. Hettiachchi, and M. Sanderson, "Can users correctly interpret machine learning explanations and simultaneously identify their limitations?" 2023, arXiv:2309.08438.
[115] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: A narrative review," *NPJ Digit. Med.*, vol. 5, no. 1, p. 46, Apr. 2022.
[116] Z. Zhang, Y. Genc, D. Wang, M. E. Ahsen, and X. Fan, "Effect of AI explanations on human perceptions of patient-facing AI-powered healthcare systems," *J. Med. Syst.*, vol. 45, no. 6, p. 64, May 2021.
[117] Z. Caplová and P. Svábová, "IBM SPSS statistics," in *Statistics and Probability in Forensic Anthropology*, Z. Obertová, A. Stewart, and C. Cattaneo, Eds., New York, NY, USA: Academic, 2020, ch. 7.1, pp. 343–352.



ADONIAS CAETANO DE OLIVEIRA received the Graduate and master's degrees in computer science. He is currently an Associate Professor with the Federal Institute of Ceará, Brazil. He is also a Ph.D. Researcher in biotechnology with Parnaíba Delta Federal University, Brazil. His research interests include artificial intelligence, digital phenotyping, explainable artificial intelligence, and the evaluation of trust in AI-based systems to improve healthcare outcomes in mental health. Key areas of investigation include the integration of technology in healthcare delivery, particularly in the following areas of health informatics: machine/deep learning, data science, natural language processing, and explainable artificial intelligence.

JOÃO PEDRO CAVALCANTI AZEVEDO is currently pursuing the integrated bachelor's and master's degrees in computer science with the Federal University of Maranhão (UFMA). He is also pursuing the master's degree in biotechnology. He is currently a Technologist in information technology management with UNICSUL.

LÍVIA RUBACK received the B.S. degree in computer science from the Federal University of Juiz de Fora (UFJF), Brazil, in 2009, and the M.S. and Ph.D. degrees in informatics from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil, in 2013 and 2017, respectively. From 2019 to 2023, she was an Assistant Professor with the Federal Rural University of Rio de Janeiro (UFRRJ), Brazil. Since 2023, she has been an Assistant Professor with the State University of Campinas (UNICAMP), Brazil. Her research interests include artificial intelligence, social computing, fairness in machine learning, and health informatics.

RAYELE MOREIRA received the Ph.D. degree in biotechnology from the Federal University of Piauí (UFPI), in 2022, where her research focused on health biotechnology. She completed her postdoctoral research with Parnaíba Delta Federal University (UFDPar) (2022–2023), specializing in digital health technologies, with an emphasis on assessment and rehabilitation. With additional training in statistical analysis and epidemiology, her current research interests include digital health, assistive technology and motor rehabilitation, evidence-based practice, health policies, and health data analysis.

SILMAR SILVA TEIXEIRA received the Ph.D. degree in mental health from the Federal University of Rio de Janeiro (IPUB/UFRJ). He is currently a Professor in physical therapy with UFDPar. He is also a Full Professor/Researcher with the Post-Graduation Program in Biotechnology and Biomedical Sciences, Parnaíba Delta Federal University (UFDPar), where he is also the Founder and a Coordinator of the Neuroinnovation Technology Laboratory (NitLab). Nitlab is a multidisciplinary laboratory focused on innovation in health for, such as implement software solutions, mobile applications, and hardware.

ARIEL SOARES TELES (Member, IEEE) received the bachelor's, master's, and Ph.D. degrees in computer science. He is currently an Associate Professor with the Federal Institute of Maranhão, Brazil, and a Visiting Senior Lecturer with the King's College London. He is also a Researcher with expertise in digital health. His research interests include leveraging technology to improve healthcare outcomes and address health challenges through cutting-edge digital solutions. Key areas of investigation include the integration of technology in healthcare delivery, in particular, on the following topics of health informatics: machine/deep learning, data analytics, natural language processing, computer vision, mobile/ubiquitous computing, and serious games.