IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 November 2023, accepted 26 December 2023, date of publication 1 January 2024,
date of current version 5 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3348783

RESEARCH ARTICLE

Dialogue System for Early Mental Illness Detection: Toward a Digital Twin Solution

AKBOBEK ABILKAIYRKYZY[1], FEDWA LAAMARTI[1,2], MUFEED HAMDI[3], AND ABDULMOTALEB EL SADDIK[1,2], (Fellow, IEEE)

[1]Computer Vision Department, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
[2]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada
[3]Department of Psychiatry, Danat Al Emarat Hospital for Women and Children, Abu Dhabi, United Arab Emirates

Corresponding author: Akbobek Abilkaiyrkyzy (akbobek.abilkaiyrkyzy@mbzuai.ac.ae)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Mohamed bin Zayed University of Artificial Intelligence.

ABSTRACT Mental health disorder rates have increased in recent years. In this research, we aim to address the barriers of stigma, accessibility, and affordability in mental healthcare by designing and developing a dialogue system that analyses the mental status of individuals. Additionally, it gives them personalized feedback based on the severity of the mental health problem. We propose a framework based on the concept of a Digital Twin for mental health, which incorporates recent advancements in technology to assess and classify the severity of mental health problems. The chatbot framework was designed in collaboration with a clinical psychiatrist and utilizes pre-trained BERT models, fine-tuned on the E-DAIC dataset, for the detection of various severity levels. The results of this study demonstrate the potential for our method to accurately detect signs of mental health problems with 69% accuracy, and high acceptability and usability with a score of 84.75%.

INDEX TERMS Digital twin, mental health, chatbot, dialogue system, depression.

I. INTRODUCTION

Mental health disorder rates have increased in recent years since the beginning of the COVID-19 outbreak. According to the World Health Organization (WHO), depression affects more than 280 million people and is the second leading cause of disability worldwide [1]. In a comprehensive study, the researchers found that major depressive disorder and anxiety disorders rose by 28% and 26% internationally last year [2]. That is, tens of millions of additional depression and anxiety problems on top of the hundreds of millions already worldwide [3]. Depression is characterized by feelings of melancholy and lack of interest. People may feel regularly anxious, low, worthless, and loss of interest. However, when this emotion persists for an extended period of time, it may become a more severe clinical case.

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

Although the number of statistics was high before the pandemic, the measures were not sufficient for several reasons. In most countries, including some high-income countries, the mental health system is not given enough facilities, resources, and awareness. Several factors contribute to this, including stigma towards mental health problems. In addition to this, access to mental health services and treatment continues to be a problem in all countries and cultures around the world, and the current clinical workforce is not sufficient to meet these demands. In developed countries, there are approximately 9 psychiatrists per 100,000 people, and as few as 0.1 psychiatrists per 1,000,000 in low-income countries [1].

The longer the mental issue remains within a person, the more significant health problems become: chronic pain and stress that may lead to heart attack, obesity, inability to cope with daily activities, and in extreme cases, suicide. In accordance with the Lancet Commission analysis on

mental health, mental diseases are on the increase in every nation and would cost the worldwide economy $16 trillion by 2030 [4].

Therefore, early detection, intervention, and proper care of any mental health problem, such as depression, can prevent severe forms of these conditions. Consequently, they may promote recovery, avoid recurrence, and lessen the financial and emotional cost of the condition [5]. To solve this problem, conversational agents have gained popularity in recent years, notably in psychoeducation and many therapy interventions; however, very few have been used for the comprehensive detection of problems related to mental health.

Artificial intelligence (AI) has the ability to deliver a personalized approach by taking detailed patient variations into consideration. A chatbot system employs conversational AI technology to imitate a natural language dialogue (or chat) with a user via messaging apps, websites, and mobile apps, among other means. It performs live chat operations in response to real-time user interactions using rule-based language applications. The Conversational AI platform employs a number of technologies, including natural language processing (NLP), natural language understanding (NLU), machine learning (ML), deep learning, and predictive analytics.

An innovative and unique approach to investigating cognitive well-being can be offered by mental health chatbots. Their use is growing in popularity due to their ability to provide a simple and private method of getting help with such problems. Chatbots may give guidance and support, as well as monitor reactions and progress over time and suggest recommendations. The information from mental healthcare chatbots is mainly a combination of clinical knowledge, technical advancement, and easy-to-interpret presentation. Chatbots can be programmed to analyze data that can be utilized to automate some aspects of clinical decision-making, and personalized recommendations for users. This could provide more opportunities for more accurate information recording and processing, fewer errors, and the ability to use past and present data to predict outcomes that eventually increase the overall efficiency of clinical services. Numerous studies have been conducted about the usability and promising prospects of chatbots in mental health [6]. Due to privacy, accessibility, and anonymity, chatbots are extremely useful in dealing with the delicate issue of mental health.

In our research, the diagnostic approach for mental health disorders employs the framework of the Diagnostic and Statistical Manual of Mental Disorders (DSM) [7]. The DSM, predominantly utilized in the United States but acknowledged globally, offers an exhaustive classification of mental health disorders with standardized diagnostic criteria. This standardization ensures uniformity in understanding and treating mental health issues across various clinical settings. The choice of the DSM in our study is underpinned by several key factors. The DSM's criteria are the result of extensive research and ongoing updates, reflecting the dynamic nature of psychiatric knowledge. This approach lends a higher degree of validation to the capabilities of our chatbot to provide diagnosis assistance. Moreover, the DSM's focus exclusively on mental health disorders, providing detailed criteria for each condition, aligns well with our chatbot's aim to identify and manage mental health symptoms. Finally, the DSM's emphasis on clinical significance criteria and its detailed symptom descriptions render it a precise tool for our research purposes, which involves a blend of statistical and qualitative analysis.

A thorough analysis of the current status of mental health support systems allowed us to shape the scope of this research. The main research question that serves as a foundation to design and develop the described system is as follows:

- How can natural language processing be leveraged to build a dialogue system for mental health, and specifically depression, with high usability?

Thereby we aim to leverage the power of AI in healthcare, in particular, mental well-being, and build an assistive tool for clinical professionals to provide symptomatic screening and identify the level of mental health problems (MHP), especially depression.

In this research, engineers and a clinical psychiatrist collaborated to devise a methodology that allows to capture insights into the current mental health status of individuals. The medical expertise of healthcare professionals allows us to apply a valuable and unique set of skills to understand the origins of mental health problems, thus enhancing prevention. In addition, involvement with mental health professionals minimizes clinical error. We develop a customized classification model to classify symptoms of mental health problems using recent developments in NLP, pre-trained BERT models, in order to detect depression. Moreover, we design and develop a chatbot system using an open-source, Rasa framework, that allows us to have a flow of conversations with users to classify the severity of MHP and test the framework usability with real participants. Our contribution lies in performing multi-class classification of mental illness severity, while most existing literature was focusing on detecting the presence or absence of such problems. Consequently, we have obtained a dialogue system that provides feedback based on the chatbot analysis. The proposed digital twin (DT) framework not only adds to the current body of knowledge, but also provides a versatile solution for other chatbot systems in the mental health domain.

II. LITERATURE REVIEW

A. CHATBOTS AND MENTAL HEALTH

The first academic exposure to this human-computer interaction technology began a half-century ago. ELIZA, the first chatbot, was created in the 1960s to try to simulate the reactions of a psychotherapist during a treatment session. It used pattern matching and pronoun replacement to provide

the illusion of knowledge despite the fact that it lacked built-in knowledge. Psychiatrist Kenneth Colby created PARRY in 1972 at Stanford University, a software capable of replicating the behavior of a human with schizophrenia, which was subsequently ''counseled'' multiple times by ELIZA. Since then, the technology that enabled ELIZA has improved using AI and ML, particularly deep neural networks, and become available worldwide for billions of people in many domains.

There are about 40 different types of mental health chatbots, mostly aimed at treating depression or autism. Many existing chatbots currently combine the concept of Cognitive Behavior Therapy (CBT) with activity tools to follow to improve the user's mental state, while few of them help manage thoughts and emotions through a combination of tools and techniques such as Dialectical Behavior Therapy (DBT), evidence-based CBT and guided meditation. And they seem to be effective: according to the findings of a 2017 study conducted by Stanford University School of Medicine researchers and the developers of the aptly named ''Woebot'', the bot was effective in reducing depression and perceived as empathetic among college students after only two weeks of therapy [8].

One of the widely adopted diagnosis technologies is Ada Health, an app-based symptom checker. Ada chatbot leverages AI to conduct interview-based communication, asking questions of various types on users' symptoms. Subsequently, the best match between symptoms and condition suggestions is predicted using an up-to-date medical database of diagnoses. Moreover, the range of the items includes personal data such as age, gender, smoker status, presence of high blood pressure, pregnancy, and diabetes. Some apps, such as Moodpath [9], ask users different questions three times a day for 14 days, based on diagnostic criteria for depressive disorders. An algorithm determines possible depression (screening) and assesses severity based on the indicated symptom patterns. Diagnostic app results are frequently based on algorithms or AI, which means that computers can simulate complex human cognition and actions. Wysa uses a combination of tools and techniques such as dialectical behavior therapy (DBT), evidence-based CBT, and guided meditation. Some application-oriented chatbots, such as Vivibot, assist young adult cancer survivors to navigate life following a cancer diagnosis by increasing resilience and decreasing distress, whereas others, such as MYLO, apply general self-help skills and aim toward suicide prevention.

The approaches for developing the Chatbot vary from bidirectional recurrent neural networks (BRNN) containing attention layers, and encoder-decoder attention mechanism architecture consisting of RNN with Long-Short-Term-Memory (LSTM) to Rasa Open-Source architecture framework. Taken together, these findings suggest that chatbots can improve mental health therapy as they can provide users with fast information [10], are available 24/7, and save users money on travel and phone rates. The main limitation of such web-based AI solutions is that they are not fully capable of understanding nuances of complex human language associated with severe mental health issues. The recent study by Palanica et al. [11] concludes the challenges with respect to the failure to understand or display human emotions. Along with poor adherence, data privacy is also a big concern for users of these services. Hence, these drawbacks need to be addressed in a manner of chatbots being able to classify emotions thoroughly, and developers of chatbots must make sure that data sharing does not expose users to privacy hazards.

A review paper by Ahmed et al. assessed the overall popularity of the 11 mental health applications, where the findings of their quality evaluation suggested a future potential in the field of anxiety and depression [12]. However, due to the limitations of publicly available data, no technological features of chatbots were thoroughly investigated. In turn, it raises attention to the importance of knowing what is behind the AI technologies that offer mental assistance. Since transparency is one of the essentials that would build trust between users and mental health applications.

The effect and preference of the different presentation modes now utilized by chatbots (text, spoken, or embodied as a 3D avatar) remain mostly unclear. As seen in our findings in Table 1, today's principal presentation modes range from text to 3D animation. While some groups have claimed that voice, rather than 3D avatar animation, is the primary determinant of a positive chatbot experience, this remains difficult to conclude today because no studies have compared adherence or engagement measures between chatbots with identical functionality but different modalities. Understanding the influence of presentation mode through a meta-analysis is challenging due to the diversity of variables involved. Even when focusing solely on usage patterns, gaining a comprehensive understanding remains a challenging task due to the complexity of these variables. Fitzpatrick et al. [8], Bickmore et al. [13] emphasize the impact of adequate rapport or therapeutic alliance on patient interactions. Although early alliance formation in conventional treatment is associated with better results, little is known about how patients feel supported by chatbots and how alliance develops and influences mental outcomes.

Table 2 highlights that the framework for performance evaluation is still being developed, therefore, an individual approach needs to be taken to interpret the results, usability, acceptance of virtual agents, chatbots, and other AI-powered tools in mental health.

Overall, the performance results of the reviewed studies show that there is room for improvement, including the lack of a monitoring system for patients, that would allow for analyzing their progress. Only a few studies investigated the long-term impact of chatbots on user interaction. Another potential gap is considered in the vulnerability of chatbot systems to misinterpretation of human language. Oftentimes, many studies followed the bottom-up approach in designing

TABLE 1. Overview of existing studies.

ID	Study	Year	Application	Key features
[14]	Hennemann et al. [14]	2022	Ada - symptom checker	Screening based on the user input symptoms and feedback with best match probabilities
[15]	Baker et al. [15]	2020	Babylon Triage and Diagnostic System	Recommendations based on user symptoms and urgency of actions
[8]	Fitzpatrick et al. [8]	2017	Woebot - a text-based conversational agent	Chatbot delivered CBT to improve self-reported symptoms of anxiety and depression
[16]	Philip et al. [16]	2017	Embodied Conversational Agents (ECAs) - a diagnostic system for major depressive disorders (MDD)	Symptomatic diagnosis of major depressive disorders (MDD) based on DSM-5 criteria evaluation during in-person interview
[17]	Denecke et al. [17]	2020	SERMO - chatbot integrated into a mobile app that incorporates CBT to manage mental health problems	Emotions are monitored and categorized through NLP and lexicon-based techniques, based on which recommendations are provided
[13]	Bickmore et al. [13]	2010	Relational Agent for Patient Education at Hospital Discharge	Patients with diagnosed depression evaluated chatbot's therapeutic relationship substantially higher than those who did not have depression, and many chose chatbot over medical professional for psychoeducation and convenience of use.

and developing chatbots that expands the gap between psychological and technical standpoints. Although many studies were evaluated in the final stages of implementation by clinical specialists such as psychologists and psychiatrists, very few have their professional input at the development stage. Further performance analysis is provided in Table 2 that demonstrates the evaluation considering the high acceptance and positive user experience feedback that were evaluated in some applications.

When making psychiatric diagnoses, reliability is especially important because there is no ''gold standard''. Several studies have been conducted on diagnosis chatbots, in particular, the diagnostic quality of the health application Ada, an app-based symptom checker. Three groups of participants (psychotherapists, psychology students, and laypersons) familiarize themselves with 20 case vignettes (12 cases from adulthood, 8 cases from childhood and adolescents) from well-known textbooks of psychiatry and clinical psychology entered by participants. The statistical outputs that were calculated as the Cohen kappa coefficient and the percentage of agreement showed low to moderate results. When comparing the types of vignette cases, the app reveals deficits in diagnosing childhood and adolescent mental disorders. Some limitations of the study presented by [18] are comprised of transferability to daily practice, the lack of a larger sample, and not including *no diagnosis present* samples since all the vignette cases had certain diagnoses and symptoms.

While the previous study was based on textbook diagnoses, the recent comparative study [14] aimed at testing the diagnostic performance of the same health application Ada on real-world patient cases. The precision of the agreement between the first and one of the first five diagnostic suggestions was assessed by Ada and the diagnoses were based on interviews with the therapists. Ada's accuracy in diagnosing conditions was moderate to good, varying from 51% of the cases for first condition suggestion to 69% for suggesting one accurate diagnosis among the first five. Although the study showed high Ada usability among participants, more than three-quarters of them preferred the traditional face-to-face approach over the app-based diagnostic. The authors highlight the need for larger samples in their analysis compared to other comprehensive diagnostic tools while noting the importance of treating such technologies rather than substituting them for health professionals.

Research shows that although symptom check technologies are highly adopted by most users due to their ease, their diagnostic performance varies significantly between other similar tools and was considered low to moderate at best [19]. The diagnostic precision of 23 symptom checkers was studied using 46 case vignettes of different emergency health conditions. Only one-third of the trial was able to predict the right diagnosis first among the list [20], while the study of 200 vignette cases investigated similar average performance precision in eight different well-known symptom checker tools such as Ada, Babylon, Buoy, K Health, Mediktor, Symptomate, WebMD, Your.MDw [21]. The correct diagnosis was listed in the top five suggestions for conditions by the symptom checkers in 40.8% of the cases, with Ada having the highest precision (77.5%). The average performance rates of 12 similar tools were confirmed by Ceney et al. [22] in the case of diagnosing the correct disease among the top five suggestions for conditions (51%).

The survey conducted by Bendig et al. [23] reviewed state-of-the-art chatbots in mental healthcare and evaluated the feasibility, effectiveness, and acceptance of these technologies in the psychological-psychotherapeutic field, as well as certain limitations in insufficient sample size

TABLE 2. Characteristic description of existing studies.

ID	Performance	Usability	Strengths	Limitations
[14]	51% in first condition suggestion, 69% in one of the top five condition suggestions	81.51% in SUS	Real patient data, free of charge, availability in 7 languages across 10 million users	Limited diagnostic spectrum, testing on larger samples
[15]	Comparable results with doctors in 80% of average recall, higher results in average precision 44.4% and F1-score 57.1%	–	Bayesian networks used in the core of the triage decisions make the solution more interpretable and explainable, opportunity to make informed decisions for users	Limited range of diseases to be diagnosed, the complexity of diagnosing multiple conditions
[8]	Depression symptoms were reduced by the PHQ-9 (F=6.47; P=0.01) over the study period, anxiety was considerably decreased by the GAD-7 (F1,54= 9.24; P=0.004)	High level of satisfaction by Woebot group(4.3; t48=3.99; P<0.001), high emotional awareness (3.3; t47.06=2.38; P=0.021)	Comparison groups (Woebot and informational control) to reduce bias in evaluation analysis, considerable progress in anxiety reduction in both groups	Choice of a control group, less diverse trial population
[16]	ECA psychometric properties for evaluated through Area Under the Curve (AUC) of 0.71, the sensitivity of 49, specificity of 93	High user acceptability (AES of 25.4), empathy, trust, psychometric measure's reliability Cronbach's alpha was 0.70	ECA can efficiently detect MDD symptom severity, has satisfactory performance compared to clinical interviews, high acceptance by users	Low accuracy in detecting mild, moderate MDD symptoms
[17]	Emotion classification shows 81% of accuracy	Values of 0.8 and higher were obtained compared to UEQ benchmark, showing high attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty	Available emotion recognition and tracking, regulation and recommendation tools, comprehensive usability evaluation, works offline	Limited range of emotions to be analyzed, lexicon-based approach constrains the recognition, poor performance with respect to unexpected user input
[13]	Patients with substantial depressive symptoms rated the agent considerably higher on therapeutic alliance than patients without severe depressive symptoms (6.2 vs. 5.5, before transformation), t(108)=2.02, p.05, d=0.58.	Only 24% of patients stated they would have preferred getting their discharge information from their doctor or nurse (40% were indifferent, 36% said they absolutely preferred the agent).	Clinical trial with a considerable number of participants (n=139), high performance and acceptance of empathetic agents for the hospital after treatment shown by patients	Results may not be generalized to other populations, the direction of the correlations between major depressive symptoms and other reported measures is unknown.

and high-quality assessment statistics. Therefore, it was suggested that the advancement in the successful use of chatbots for mental healthcare could be accelerated by studying clinical samples, given that it performs under a safe and guideline-based framework. Studies often follow a bottom-up approach in developing their chatbot, which addresses the problem from a technical background followed by a collaboration with clinical and psychological scientists. Whereas the evaluation of field specialists' needs first would minimize the theoretical gap to meet the expectations of psychological endpoints. There is a need to replace studies with randomized controlled trials in a clinical sample and provide higher-quality statistical power. In general, chatbots can play a clinical role in addressing people who are unaware of receiving treatment due to various barriers.

Although patients were experiencing a rather positive outlook on the adoption of AI-based mental health tools, healthcare providers appear to be more skeptical. This evidence could be supported by relatively lower diagnostic performance compared to the professionals' diagnosis. However, it seems to converge when more condition suggestions were considered in predicting the diagnosis. According to Semigran's research, clinical vignettes were diagnosed correctly among 84.3% of the top three diagnoses by physicians, while only 51.2% of symptom checkers could achieve that [24]. However, the inclusion of five diagnostic suggestions revealed that Ada shows similar diagnostic performance compared to healthcare professionals (77.5% and 82.8%, respectively). Babylon Diagnostic and physicians' diagnosis achieved comparable diagnostic performance while estimating the diagnosis among the top five conditions [15].

In addition to the diagnostic performance of existing tools, usability, and user experience were investigated in several studies. The usability of the iHelpr chatbot, mental health care within a social enterprise, was assessed through the USE questionnaire [25], Software Usability Measurement Inventory (SUMI), the System Usability Scale (SUS) [26]. The SUS score for iHelpr Chatbot was 88.2, where the average result was above the 68 scores. Chatbottest has developed a collaborative guide of questions that fall under seven different categories to test the specific functionality of chatbots: Answering, Error management, Intelligence, Navigation, Onboarding, Personality, and Understanding.

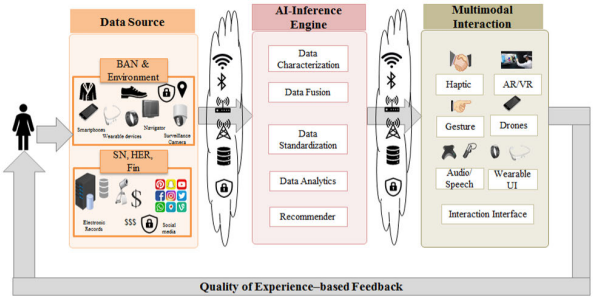

FIGURE 1. Digital Twin Ecosystem for health and well-being [29].

The average percentage for the iHelpr Chatbot was 55.6%. The lowest result was 43% and the highest was 74%.

Many review articles reveal that existing applications are primarily concerned with addressing mental health issues, with very few focused on diagnosing them. Some studies have been conducted on recommendation tools for a matching system between doctors and patients. However, few works are intended to provide individual recommendations to improve an individual's mental health. In fact, we notice that treatments are based on preliminary, defined schemes. Since depression or anxiety levels can vary between patients, it is important to address these individual symptoms instead of following a predefined path. The PHQ-9 questionnaire was chosen as a ground truth based on the recommendations of a psychiatrist and a review of the existing literature [27], which objects to and evaluates the severity of depression through a questionnaire.

It is worth mentioning that the results of some works should be interpreted with proper care since there is a high probability of conflict of interest. Although much work on the potential of chatbots in mental healthcare has been carried out, there are still some critical issues with privacy, the ability to adapt due to a limited dataset, and trials being performed primarily for a broader audience and in different languages. Also, as it was highlighted several times, the main attribute here is to acknowledge that chatbots initially are not intended to fully replace psychology specialists but rather act as complementary tools. This understanding will open up more space for creativity, and breakthroughs that eventually will work for the sake of humanity.

B. DIGITAL TWINS

The fundamental role in making healthcare more accessible, and up-to-date, is being played by emerging digital technologies. Digital twins have received much attention in the last decade due to significant improvements in the manufacturing sector. According to Gartner [28], "The implementation of a digital twin is an encapsulated software object or model that mirrors a unique physical object, process, organization, person or other abstraction". Therefore, digital twins can represent not only physical objects, but also processes or other abstract concepts. However, the potential of DT was not only limited to these areas, as more recent evidence highlights the successful application of DT in healthcare. The health and well-being of the population can benefit from incorporating DT in their daily life by increasing awareness about their mental health status by means of feedback and providing individual recommendations.

According to El Saddik [30], a digital twin enables the monitoring, understanding, and optimization of the physical entity's functioning, as well as providing constant feedback to improve quality of life and well-being. Machine learning techniques can be leveraged to capture lifestyle trends and anticipate prospective health issues [30]. Furthermore, contextual data, emotional state, and preferences might be obtained and processed in order to completely grasp and define the person's holistic condition.

Personal health system requirements along with DT characteristics are used to construct the ecosystem of DT for healthcare [29]. The key requirements for designing a personal health system proposed by Badawi et al. [31] include interoperability, usability, personalization, feedback, mobility, accessibility, and security. Fig. 1 presents the pipeline with the primary purpose to provide the foundation for a sustainable DT system in preventative healthcare. According to this set of characteristics, preventive health DT must use sensors (hard and soft sensors) to capture health data, actuators to provide feedback, and an AI core component to perform data analytics and make recommendations [30]. Over time, the health data is collected in structured storage in the cloud, containing the information related to the physical twin, and providing highly useful information on the health and well-being condition of individuals [32].

As DT aims at improving the well-being and quality of life of individuals, it also involves identifying the emotional state through monitoring and analyzing using machine learning, to provide Personalized Healthcare [33]. Another key current element of a digital twin is its capacity to predict the behavior or the outcome of the process, and the accuracy of the predictions is continuously improving due to integration with artificial intelligence and sophisticated analytics. As a result of this feature, the analyzed results are communicated as part of the feedback loop to the original physical object. A recent review on this subject stated that precision public health will play a key role in providing tailored healthcare and intervention techniques [34]. Authors describe precision mental health as a preventative and intervention strategy that pays special attention to initial assessment, continuous monitoring, and tailored feedback information to get an accurate picture of an individual's requirements, preferences, and prognosis possibilities, and then tailor therapies and support appropriately in accordance with the most recent scientific data. Some studies claim that there must be a structured approach that employs valid, reliable, and standardized metrics of feedback systems in mental health interventions such as Measurement Feedback System (MFS) [35]. The efficacy of such an approach is shown in the demonstration of MFS to enhance results in adult

mental health, particularly for individuals who were neither improving nor worsening while in treatment. Other valuable tools, such as routine outcome monitoring (ROM) and clinical feedback systems (CFS), for averting treatment failure and improving therapeutic results, particularly in patients who are not advancing in therapy, are becoming more common in mental healthcare. There is still considerable ambiguity with the structure and quality of the feedback provided in mental healthcare. Therefore, in this study, we developed the feedback component based on the recommendations of a clinical psychiatrist and individual symptom characteristics.

This enables us to build a history and track changes in the emotional background of a person even though it is not a complete digital twin system. The digital twin employs hard and soft sensors to continually gather data and generate an accurate reproduction of the physical twin at any given time. These sensors can take the shape of wearables and personal health gadgets. By connecting wearable and IoT devices to our mental health system, we can collect physiological and contextual data that subsequently helps to find a relationship between life events and the corresponding mental state of an individual. Therefore, a dialogue system helps to contribute to building mental state DT, allowing us to get information from the user and store the history in the cloud system.

C. DIGITAL TWINS IN MENTAL HEALTH

In recent developments within the field of mental health, the concept of a ''digital twin'' has transcended its traditional association with physical object modeling to encompass the virtual representation of an individual's mental state.

Our research aligns with and extends the innovative perspective of digital twins in mental healthcare, as seen in the following key studies. Spitzer et al. [36] effectively illustrate this transition by exploring how digital twins can mirror individual mental states and responses, thereby contributing to precision mental health. This indicates a paradigm shift where the digital twin concept extends beyond the representation of physical entities to encompass the dynamic and abstract nature of human cognition and mental health.

Similarly, the authors of [37] provide further research in this expanded scope by showing how digital twins can monitor and enhance well-being, a concept that inherently involves abstract mental and emotional states. This broader application underlines the adaptability of digital twins to represent not just physical objects but also intangible, conceptual elements.

The study [38] provides a practical viewpoint, focusing on the technical challenges and solutions in creating an effective digital twin for mental wellbeing, using machine learning. Lastly, Miller et al. [39] delve into the concept of a digital twin expanded beyond representing physical entities, illustrating its application in holistically simulating and predicting an individual's well-being, inclusive of mental, physical, and financial aspects. Collectively, these sources underscore the evolving scope of digital twins in healthcare, emphasizing their significant potential in personalized and precision mental health interventions.

Our work leverages this concept by creating a dynamic, AI-driven model that mirrors and analyzes an individual's mental state. This approach is not only in line with the latest advancements in digital health technologies but also broadens the application of digital twins to include mental health monitoring and early illness detection. Therefore, the term ''digital twin'' in our research describes the virtual representation of mental states, with the aim of offering an assistive tool in personalized mental healthcare and intervention.

In this study, our objective was to contribute to the field of mental health technology by developing a dialogue system enhanced with conversational AI. This system is designed to detect mental health issues and provide personalized feedback and recommendations, based on each individual's mental health profile analyzed by our chatbot. Our approach combines AI with domain expertise to address mental health detection challenges, and from a clinical perspective, it could potentially aid in reducing stigma and increasing accessibility in mental health services. In our experimental study, the system showed promising usability and accuracy.

A key aspect of our research is the advancement in classifying the severity of mental health problems, which extends beyond the typical binary detection focus of existing studies. The dialogue system we developed also provides feedback based on the chatbot's analysis, a feature not commonly emphasized in current chatbot solutions. Additionally, we proposed a digital twin framework in mental health, with the potential of offering assistance to professionals in personalized mental health care. This study provides a performance evaluation on both existing datasets and with real participants. Our findings suggest that recent advancements in natural language processing could be beneficial in supporting mental health professionals and improving self-awareness about mental health conditions.

III. METHODOLOGY

A. PROPOSED FRAMEWORK FOR MENTAL HEALTH DIGITAL TWIN

Chatbots are among the most accessible tools that leverage AI through natural language communication for humans. Often, they represent medical expertise, AI developments, and a convenient user experience in one system. Chatbot, as an important use of NLP, enables computers to grasp the meaning of natural language and respond appropriately. The latest natural language understanding of chatbots is based on deep learning frameworks due to better performance in most NLP tasks.

The algorithm integrated into our chatbot framework for mental health assessment has been collaboratively developed with input from clinical psychiatrists. Our primary objective in this collaboration is to create a tool that can assist clinicians

in their work. To ensure an effective conversation flow for the chatbot, we have drawn inspiration from standard clinical interviews. These interviews typically begin by inquiring about recent issues individuals may be facing, encompassing both physical and mental health concerns. Once a specific problem is identified, we pose targeted questions related to mental health to assess its severity. Our chatbot encompasses seven key life domains that individuals may discuss with a therapist or counselor. These domains, including well-being, physical health, choice and control, hope and hopelessness, self-perception, relationships, support systems, and activity levels, collectively define the overall quality of life.

It is important to clarify within our study, that the legal responsibility rests with the individual user. Our system is designed to provide informational support rather than direct intervention. This approach is aligned with the objective of the chatbot to serve as an informational and preliminary screening tool, directing users to appropriate professional resources in times of acute mental health crises.

In this research, we are collecting data on the person's mental state through the chatbot system. To address concerns regarding privacy, we implemented a set of measures to ensure the protection of participants' data throughout our research process. Firstly, we employed anonymization techniques to dissociate any personally identifiable information from the collected data. Each participant was assigned a unique ID, and all data were saved using these IDs rather than their actual names or personal details. Furthermore, we prioritized secure data storage practices. The collected data were stored on dedicated hardware infrastructure that featured multiple layers of password protection. Only authorized researchers and supervisors were granted access to this hardware, and stringent access control mechanisms were in place to safeguard against unauthorized access or breaches.

Our study on the incorporation of digital twins in mental health takes privacy into consideration. The interaction between the web application and the user ensures privacy through various measures. Firstly, we adhere to relevant data protection to ensure the confidentiality and security of user data. Our application incorporates privacy-enhancing features, including access controls, and data retention periods limited to users' preferences. The access controls enable users to have control over their data, empowering them to make informed choices regarding information sharing. Additionally, our authentication methods, login procedures, secure data transfer over hypertext transfer protocol secure (HTTPS), and unique user identifiers contribute to the secure identification of users within the web application. Through these measures, we prioritize user privacy and strive to create a safe environment for the application of digital twins in mental health.

The study by Schrank et al. [40] highlights the importance of insight into mental health disorders and its impact on treatment adherence. Insight refers to the awareness of a mental disorder, its consequences, the need for treatment, symptoms, and attributing symptoms to the disorder [41]. It has been found that higher levels of insight are associated with positive clinical variables, including better treatment adherence. Drawing from this understanding, our dialogue system for mental health problem detection aims to address the issue of insight in depression patients. By engaging in conversations with users, the chatbot system plays an important role in increasing awareness and understanding of mental health disorders, including depression. It helps individuals recognize the symptoms they are experiencing and attribute them to a mental health condition. Moreover, the system emphasizes the potential consequences of untreated mental health problems, highlighting the importance of seeking appropriate treatment.

Through this approach, the dialogue system seeks to enhance users' insight into their mental health condition, promoting self-awareness and encouraging them to take proactive steps towards seeking professional help and adhering to treatment plans. By facilitating discussions and providing information, the system contributes to the overall goal of supporting individuals in recognizing and addressing their mental health needs.

In this paper, we propose a framework that leverages the concept of Digital Twins (DT) for mental health monitoring and intervention. Our work integrates four interconnected components: Data Collection, Data Processing & Model Training, NLP Processing, and Core, tied together with a Feedback Loop, all of which form the essence of a mental health AI dialogue system, as depicted in Fig. 2.

In the Data Collection module, we acquire data, such as user profile, dialogue data, PHQ scores, etc. Then in the Data Processing & Model Training phase, we process the dialogue data through several stages. On these data, we performed data preprocessing to ensure data quality, proper data formatting, data labeling following the PHQ scores used as ground truth, data augmentation, etc. These steps are explained in detail later. The custom model is then trained, tested, and evaluated for accuracy, laying the foundation for an effective intent classification in the later stages.

When a user interacts with the chatbot in real-time, as the dialogue progresses through the Chatbot interface, the user input undergoes tokenization by the BERT Tokenizer and feature extraction via the Language Model Featurizer, leading to an Intent Classification stage, as we explain in more detail in the next section. This stage features a dual approach with a Custom Classifier and a Fallback Classifier to guarantee the accurate interpretation of the user's intent.

The core of our framework is anchored by the Analysis phase, where it classifies user intents through dialogue, continuously reiterating and averaging this process for a more accurate understanding of the user's mental state. Following this, the Mental Health Status Assessment phase finalizes the user's mental state into "No Symptoms," "Mild," or "Moderate or Severe," based on the initial analysis. Based on the results of this step, the Feedback Loop offers a means of interaction with the user, providing the next chatbot dialogue sequence. It also provides personalized recommendations

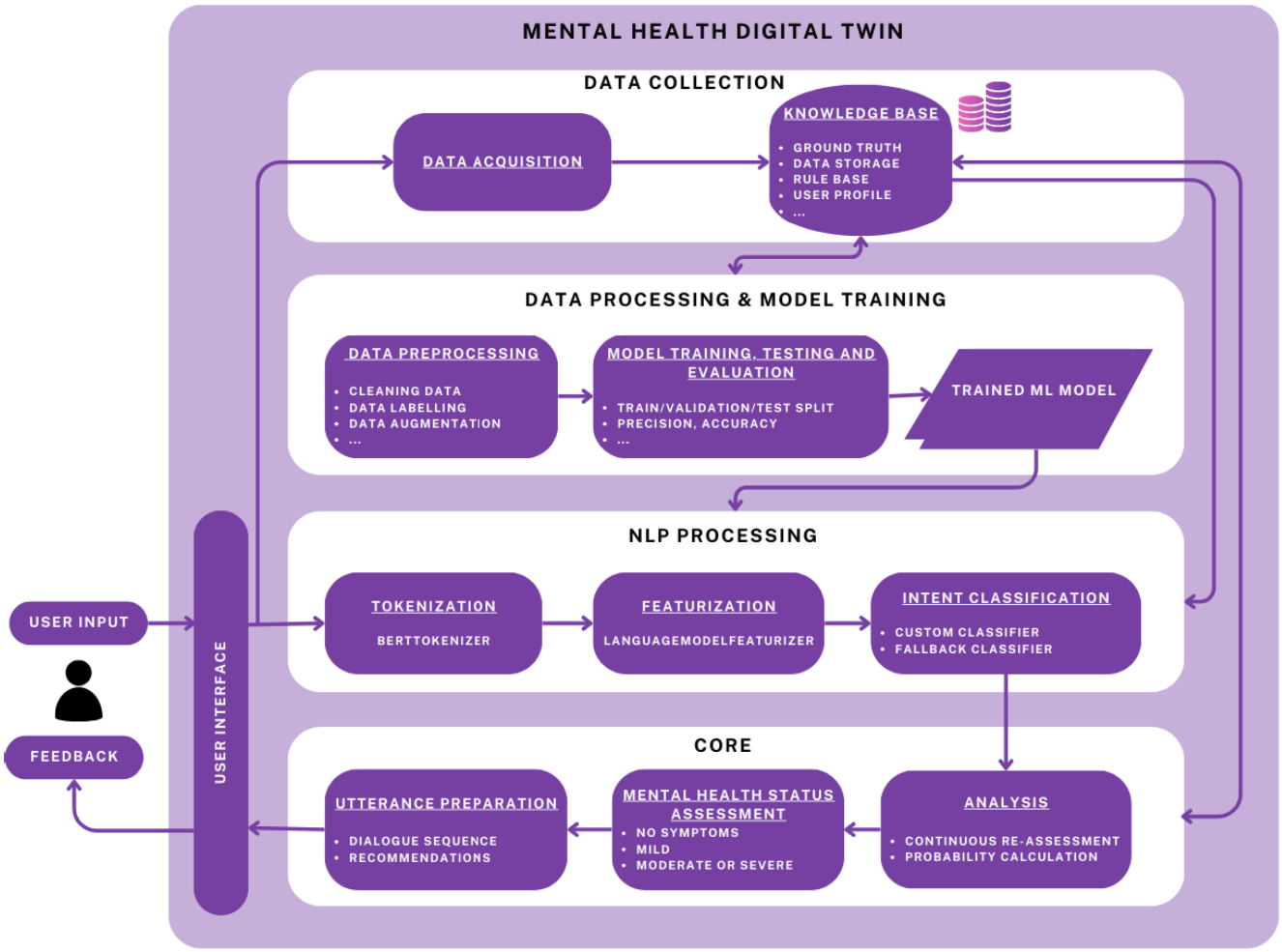

FIGURE 2. Proposed mental health digital twin framework.

designed in collaboration with the clinical psychiatrist. The objective is to improve users' self-awareness and management of their mental health. For some users who agree to take the PHQ test periodically, we can keep improving the performance of the chatbot by labelling the user dialogue with the appropriate mental health assessment depending on the PHQ score. This information will be fed back to the model training to keep enhancing the accuracy of the predictions over time. Ultimately, our framework is designed as an adaptive system, capable of learning and evolving, to provide a supportive Digital Twin for mental health care.

B. CHATBOT DESIGN

Rasa open-source Python framework was used in this study, to develop a conversational AI-powered emotional mental intelligence (EMI) chatbot. The purpose of the chatbot framework is to ensure sufficient conversation flow. Rasa Open-Source architecture consists of two main models: a Natural Language Understanding (NLU) model and a core model. The Rasa NLU is built to recognize the intents of the user's utterances and to extract entities from them. The Rasa core model is responsible for how well the conversation flows: memorizing the entities, generating a meaningful and proper response and/or action, and understanding the order of subsequent utterances coming from the user. It is feasible to train NLU and core models individually using Rasa on separate datasets; however, we need both components to create a full agent capable of interacting with a user.

There are several stages involved in this research: data collection and pre-processing, chatbot design and analysis, development, web deployment, and testing. The Rasa workflow diagram inspired by [42] gives us an in-depth look at the process of natural language modeling in the Rasa framework. Rasa NLU is an open-source natural language understanding module that provides a classification of user intents and entity extraction. This model is comprised of NLP and ML libraries and models. While Rasa Core, as it was mentioned earlier, coordinates conversations according to scenarios for chatbot-user interactions. The workflow begins with the simulation of training data. In particular, user intents and examples can be integrated into the NLU data file. Stories for different scenarios (happy path, unhappy path, etc.) are

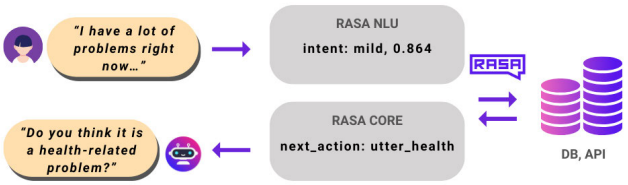

FIGURE 3. Rasa architecture.

available in designated locations. Another type of data used to train the dialogue management model, rules, contains certain conversations that are meant to follow the predefined scenario regardless of any circumstances. More information will be given in the following sections. Once the data is inserted, the dialogue interaction is constructed by grouping user intents, chatbot utterances, and action items in a logical manner. The training of the NLU and core models is performed to enable intent matching and entity recognition across the chatbot workflow. Once trained, the chatbot performance is evaluated through the means of evaluation metrics.

The Rasa processing pipeline is a main component of the Rasa NLU model that allows for the efficient processing of unstructured data. It outlines the various stages that incoming user messages must pass through in order to generate a model output. The chatbot's dialogue management flow, as depicted in Fig. 3, demonstrates how the NLU model of the Rasa framework translates input messages into a dictionary containing the original text and the intent.

The Rasa NLU framework offers a wide range of pre-defined components that allow for the customization of the chatbot model. Additionally, custom components can be implemented by incorporating pre-trained models for specific downstream tasks or by utilizing NLU training data and continually improving it with additional collected data.

Tokenizers, featurizers, intent classifiers, and entity extractors are the main components of the chatbot pipeline, as explained in the following subsections:

1) TOKENIZER

Tokenization is the process of extracting a list of words (tokens) from a spoken phrase. Instead of sentences, many features are derived from words. Tokens generated by this phase will be applied later in the pipeline for extracting features. We used `BertTokenizer` since it is a crucial component in using the pre-trained BERT model [43]. It is responsible for converting text input into a numerical representation that can be processed by the BERT model. The tokenizer has a fixed vocabulary and a specific way of handling out-of-vocabulary words, ensuring the consistency of inputs to the model.

2) FEATURIZER

Featurization is the process of converting words into meaningful numbers (or vectors) that can be given to the training algorithm. In our pipeline, we included the type of featurizer that captures dense features. A vector representation of user input is computed by `LanguageModelFeaturizer` using a pre-trained model [44].

Pre-trained language models, such as BERT, have generated a significant amount of interest in recent years. Although they may provide outstanding results on NLP tasks, they are also resource-intensive. Many people wonder whether the advantages of adopting big pre-trained models balance the increases in training time and computational resources. Therefore, we used several existing pre-trained language models from the Huggingface platform to compare their performance to the baseline BERT model [45], which is a self-supervised transformer model that was pre-trained on a huge corpus of English data.

3) INTENT CLASSIFIER

Once all the features have been generated, the intent classification model can be utilized. There are several built-in intent classifiers, such as MitieIntentClassifier, LogisticRegressionClassifier, SklearnIntentClassifier, KeywordIntentClassifier, DIETClassifier, and FallbackClassifier within the Rasa framework. For more accurate performance, we constructed a customized intent classification model by fine-tuning BERT. Pre-trained BERT tokenizer was used to convert the text in our data to token IDs. We then mapped each of the clusters ('moderate or severe', 'mild', and 'no symptoms') to classes (0,1,2), respectively. The data was split into 85% for training and 15% for validation. BERT was instantiated for the sequence classification model and created training and validation dataloaders using batch sizes of 4 and 32 respectively.

BERT takes the sentences in our data and converts them into tokens, which are then transformed into contextualized embeddings. The output is then passed to a fully connected layer of size three, where each node represents one of the classes we are trying to predict. The BERT sequence classification model is then fine-tuned to our dataset. AdamW was initialized as our optimizer with a learning rate of $1e^{-5}$ and an epsilon of $1e^{-8}$. Furthermore, we used a linear warm-up scheduler for our optimizer. Softmax is applied to allow us to interpret the output of the final layer as a probability distribution over the multiple classes.

Another intent classification model that we used in the chatbot pipeline is `Fallback Classifier`, which handles out-of-scope user intents. If the previous intent classifier was unable to classify the user message according to the set threshold (classification confidence should be greater or equal to this value), then it will be classified as nlu_fallback intent. In such cases, we use Fallback Action to handle unclear NLU predictions by asking users to rephrase their message. Every time this intent is detected, the user receives the same default fallback response, which might give an impression of a robotic manner. Therefore, we developed a custom action class that allows the NLU model to provide many alternatives and prevent repetitive fallback prompts.

4) POLICIES

The policies that our assistant uses from a library and follows the specific order to determine which action to take at each stage of a conversation are listed below based:

- `MemoizationPolicy` recalls stories from the training data and checks the correspondence of current dialogue to the stories data file. If so, it will predict the next action with a confidence of 1.0 based on the stories in our training data. If no matching conversations are detected, the policy predicts None with a confidence level of 0.0.
- `Transformer Embedding Dialogue (TED) Policy` is a multi-task transformer architecture that is used to improve the performance of dialogue systems by providing a more accurate representation of the dialogue context, which can help the system make better decisions about the next action to take.
- `RulePolicy` is responsible for dialog sections that have consistent behavior, where the predictions are generated according to any predefined rules.

IV. EXPERIMENTS AND RESULTS

A. DIALOGUE SYSTEM DATASET

The Extended Distress Analysis Interview Corpus (E-DAIC) [46] served as the foundational dataset for training our chatbot. This dataset includes semiclinical interviews designed to facilitate the identification of psychological distress disorders such as depression and post-traumatic stress disorder (PTSD). The dataset is multifaceted, featuring audio and video recordings, as well as questionnaire responses. Within this corpus, we focus on the Wizard-of-Oz interviews among 219 participants, conducted by an animated virtual interviewer named Ellie, controlled by a real human interviewer from a separate room. A range of verbal and nonverbal elements have been transcribed and annotated in the data. The dataset, excluding the unreleased test set, is divided into a training set of 163 samples and a development set of 56 samples, but the general variety of the speakers is retained in terms of age, gender, interviews, and eight-item Patient Health Questionnaire scores (PHQ-8) [47]. The PHQ-8 aims to measure the severity of depression and the scale is 0 to 24 with cutpoints at 5, 10, 15, and 20 for categorizing mild, moderate, moderately severe, and severe depression. The E-DAIC contains a transcript of the interactions that were automatically transcribed using Google Cloud's speech recognition service, audio files from the participants, and their facial features. The initial training data consists of participant IDs, binary labels (PHQ-8 scores 10), actual PHQ-8 scores, participant gender, and detailed responses to all questions in the PHQ-8 questionnaire. Additionally, we have integrated interview content into training and validation sets, encompassing various aspects such as personal traits, social interactions, life regrets, mental health challenges, coping mechanisms, previous diagnoses, and traumatic experiences. For instance, excerpts like "I wish I wouldn't have invested into…" and "so…sleeping is just not really something I do well" provide insights into the dataset's extensive content [46].

TABLE 3. Distribution of PHQ scores according to depression diagnostic status [27] and interpretation of severity levels in our model.

PHQ-score range	Questionnaire	Our model
0-4	No symptoms	No symptoms
5-9	Mild	Mild
10-14	Moderate	Moderate or Severe
15-19	Moderate to severe	
20-27	Severe	

For our studies, we selected the PHQ-9 [27] despite using a training dataset based on the PHQ-8 for several reasons. Firstly, numerous studies have indicated that the PHQ-8 and PHQ-9 demonstrate comparable performance in screening for major depressive disorder (MDD) [48]. In line with this research, a comparison study by Shin et al. [49] noted that the PHQ-8 was as useful as the PHQ-9 for MDD screening in a psychiatric outpatient sample. Therefore, our decision to experiment with PHQ-9 aligns with existing evidence suggesting an equivalence between the two versions. Furthermore, our collaboration with a psychiatrist resulted in the endorsement of the PHQ-9 as an appropriate screening tool for depression.

1) DATA PREPROCESSING

Several data preprocessing techniques were performed on the E-DAIC dataset in order to turn raw-structured data into the rasa-readable format. First, the identification number (ID), interview transcripts, and PHQ-8 scores of participants were collected from the dataset. According to the validation of PHQ scores to measure the severity of depression symptoms, data samples with more than 50 character lengths were distributed into 5 classes, where cut-off points are shown in Table 3.

After data samples were extracted in the required format, data imbalance was observed between several classes. Data augmentation might increase classification model performance, particularly for the minority class, perhaps extending a new solution to the data imbalance issue. Among many existing methods in the literature, Fadaee et al. [50] substitute target words with contextualized word embeddings rather than static word embeddings. In Data Augmentation for Low-Resource Neural Machine Translation, they employ text augmentation to evaluate the machine translation model. In his study of contextual augmentation, Kobayashi [51] advocated using a bi-directional language model. Following the selection of the target word, the model will suggest suitable replacements by providing surrounding terms. Bi-directional architecture is used to learn both rightward and leftward context since the target might appear in any location of the phrase.

2) TRAINING DATA

This section describes various types and structures of training data that were used to train Rasa assistants. Rasa Open Source

manages all training data, including NLU data, stories, and rules, using YAML as a consistent and flexible language.

- NLU training data includes classified examples of user utterances. Entities may also be used as training examples, as keywords taken from a user's message. We used several classes of depression symptoms from the E-DAIC dataset and distributed them into intent categories.
- Stories file contains the possible scenario that controls the flow of conversation between the user and chatbot. It is used to train the dialogue management model to recognize patterns in conversations and generalize to previously unseen data.
- Rules are used to train the RulePolicy by describing short chunks of discussions that should always follow the same order.
- Domain covers the list of intents, actions, utterances, forms, and slots. Each utterance demonstrates the chatbot's response to user intents that are based on the seven domains of quality life that can help demonstrate concern for someone in difficult circumstances, elaborate on their physical health status, self-perception, independence, choice and control, hope, belonging, and relationships [52].
- Test stories are used to test the performance of a trained dialogue model. These stories are written in a specific format, which is used to simulate a conversation and test the model's ability to handle different types of input and respond appropriately.
- Custom actions were incorporated with our classification model based on fine-tuning pre-trained BERT.

B. EVALUATION METRICS

It is essential to evaluate model performance using multiple metrics, as it ensures the correct operation and optimization of the model [53]. Performance evaluation is a procedure to assess the classification capacity of a system or model. According to Dalianis et al. [53], it is advised to use confusion matrix [54], accuracy [55], precision [56], recall [57], and F1 score [58] among many other existing evaluation metrics.

The following equations (1), (2), (3), and (4) were used to calculate our evaluation metrics, where TP stands for True Positives, FP for False Positives, TN for True Negatives, and FN for False Negatives:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$F1\ Score = \frac{2 \times (Precision) \times (Recall)}{Precision + Recall} \tag{4}$$

Our experiments are divided into two sections: evaluation using the E-DAIC dataset and experiment with real participants.

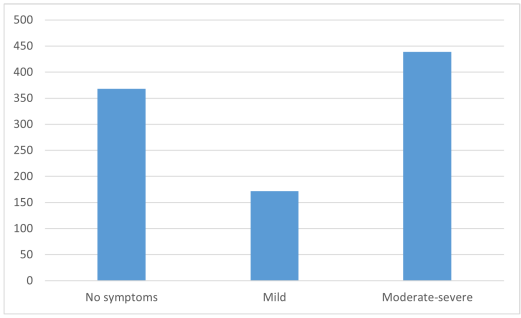

FIGURE 4. Distribution of dataset after pre-processing.

TABLE 4. Classification report.

Class Name	Label	Accuracy	F1	Precision	Recall
Mild	0	0.46	0.43	0.40	0.46
Moderate or Severe	1	0.77	0.77	0.76	0.77
No symptoms	2	0.70	0.73	0.76	0.70
Accuracy			0.69		
Macro Average			0.64	0.64	0.64
Weighted Average			0.70	0.69	0.69

C. TESTING ON E-DAIC DATASET

First, we evaluate the performance of mental health problem level classification with the pre-processed E-DAIC dataset using our chatbot. In data class distribution, the issue of slight data imbalance and lack of enough data samples remain one of the main challenges. Thus, the combination of several classes showed improved performance. Ultimately, our Emi chatbot classifies mental health problems into three classes: no symptoms, mild, and moderate-severe, as we can see in Fig. 4. The evaluation metrics of F1 score, accuracy, precision, recall, and support were used for each class during the multi-label classification of levels for depression symptoms using the Emi chatbot system, as shown in Table 4. The combination of classes positively influenced the performance of the intent classification model.

The accuracy of this model is 69%, which is higher performance than a similar mental health-related system, the Ada symptom checker, where the prediction of accurate diagnoses was the first among the top five results 51%. However, we should consider that the range of diagnoses that will be screened by Ada is not limited to only mental health-related issues. Whereas, our Emi chatbot aims to investigate the status of mental health-related issues, which by nature are known to be more complex.

Acknowledging the limited number of effective models currently available, our study represents a step forward, showing comparable results to existing benchmarks. While there is significant potential for progress, our initial outcomes are promising. They reflect a commitment to enhancing AI dialogue systems as adjunct tools for mental health care. Our framework serves as a preliminary yet solid foundation for ongoing development, aiming to incrementally refine the role of AI in mental health domain. We are dedicated to continual improvement and see this research as an initial stride toward better support for mental health professionals.

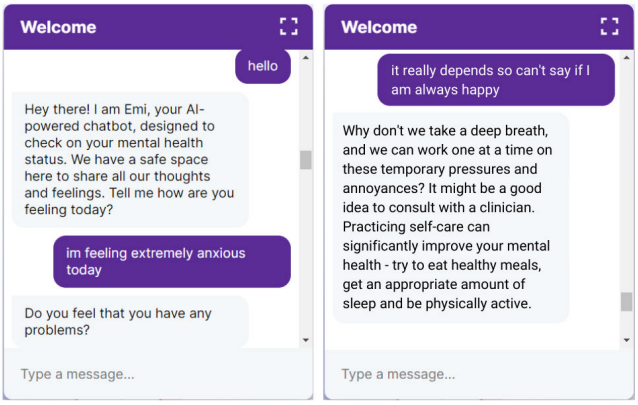

FIGURE 5. Example of user conversation with Emi chatbot (a), and Emi chatbot encouraging a user to seek professional help (b).

D. TESTING WITH REAL PARTICIPANTS

1) DEPLOYMENT

We developed a front-end page for the purpose of conducting experiments with real participants, where they can follow steps and interact with the chatbot. This frontend page is developed using HTML, CSS, Bootstrap, and JavaScript. The backend is developed with Flask. Testers utilize the Web application to provide feedback and qualitative assessment. Rasa framework allows integrating the frontend. For that purpose rasa-webchat which provides easy integration of virtual assistants with a website is used. Users can have a conversation with our Emi chatbot through this chat widget as shown in Fig. 5.

Before the users start a conversation with the mental health chatbot, we make sure to make them aware of its domain and purpose, in helping them gain an understanding of their mental health. The chatbot displays its introduction message, saying, 'Hey there! I am Emi, your AI-powered chatbot, designed to check on your mental health status' when users initiate a conversation. Additionally, we had a small conversation with the participants on the purpose of the study in helping with mental health, explaining that the main aim of the chatbot is to assess their current mental health condition. Therefore, participants are encouraged to have an open and transparent conversation to facilitate accurate assessment.

This explanation lets interested users be motivated to provide information to the chatbot as they actively seek knowledge and resources to enhance their well-being. By engaging with the chatbot, users demonstrate a proactive approach to their mental health, indicating a motivation to gain insights, learn coping strategies, and explore avenues for self-improvement. The chatbot serves as a valuable tool in their journey toward mental wellness, providing a convenient and accessible platform for obtaining information and support.

Therefore, users will be motivated to provide information to the chatbot for several reasons. Firstly, sharing relevant information enables the chatbot to better understand their unique situation and provide more accurate and personalized support. By disclosing their experiences, symptoms, and concerns, users can receive targeted advice and guidance tailored to their specific needs. Secondly, providing information to the chatbot facilitates early detection of depression symptoms. The chatbot's algorithms can identify patterns, assess risk factors, and offer timely interventions, potentially preventing the escalation of mental health conditions. Lastly, sharing information with the chatbot grants users guidance toward professional assistance. The chatbot can ensure they receive appropriate guidance. Overall, by actively engaging with the chatbot and sharing information, users can enhance their mental well-being and receive the necessary support they require.

2) CHATBOT ANALYSIS

As we discussed earlier, the E-DAIC dataset is limited to clinical interviews. Addressing such complex task related to mental health requires us to collect more data on an individual's background. Therefore, to further evaluate and improve the Emi chatbot, we decided to conduct experiments with a number of participants that can provide conversational data. We recruited 20 participants among university students (n=20, 12 males and 8 females), who were 21 years old and above with/without previous diagnosis of mental health problems. After accessing our web application, certain procedures were followed by each participant for this part of the experiment:

- The experiment procedure is explained to the research participant.
- The informed consent form is signed by the research participant.
- Screening tools as psychological questionnaires (PHQ-9, PCL-C) are taken by a research participant.
- Conversation between Emi chatbot and the participant happens.
- User experience feedback is provided by the research participant.

The PHQ and PCL-C questionnaire scores are used as ground truth to compare the efficacy and performance of the Emi chatbot component.

According to the PHQ-9 screening results, 40% and 30% of the participants were mildly and moderately depressed respectively, while 20% had no symptoms of depression. PCL-C questionnaire revealed only 10 to 20% of participants having some moderate symptoms. According to the chatbot analysis result, accuracy performance was 40% on classifying the depression severity. However, we achieve 65% accuracy when detecting mental health problem classification correctly among the top two classes. While having an open conversation with EMI, most of the users expressed quite positive feedback on the Chatbot's entertaining capabilities (sharing joyful pictures of animals, memes, etc.). However, this led us to acknowledge a system limitation. When interacting with the Chatbot, some participants actually felt a connection

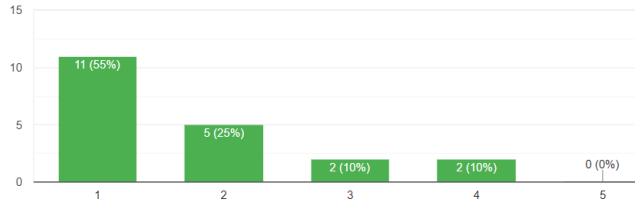

FIGURE 6. Distribution of responses for the prompt of necessity to learn a lot of things before using the system, where "Strongly disagree" is 1 and "Strongly agree" is 5.

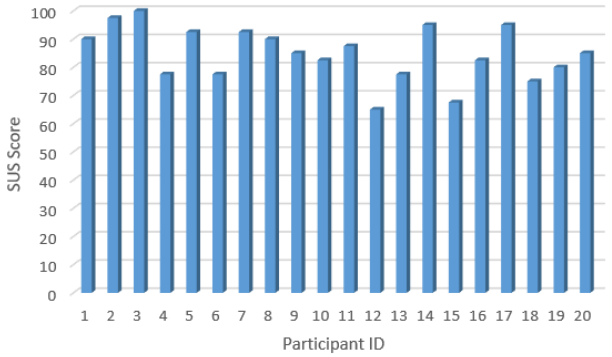

FIGURE 7. System Usability Scale (SUS) score results.

with it: when prompted with a cheerful response, they would spontaneously react with a large smile or a vocal expression indicating contentment, before replying in written form. This visual response was obviously not captured by the Chatbot, which makes it a missing key component that could have helped to more accurately predict the users' mental status.

3) USABILITY ANALYSIS

The usability and acceptability of our system were also evaluated by participants using System Usability Scale [26], where the main prompts analyzed the complexity, learning rate, need for assistance, integration, and inconsistency, among others. Scales of 1-5 describe the user's agreement with the given statement, where 'Strongly disagree' is 1 and 'Strongly agree' is 5 points. We found that 80% of the respondents do not think that extensive learning is needed before using this system as can be shown in Fig. 6. No participants found the system complex to use and they agreed that it is easy to learn for most users in a short period of time. In general, a high acceptance and usability rate can be observed in the evaluation, as can be shown in Fig. 7.

Further interpretation of the overall SUS score was performed with the calculation of each participant's SUS score in accordance with guidelines given by Brooke et al. According to the results of the usability test in Fig. 7, the Emi chatbot received 84.75% of the average SUS score, indicating higher results compared to the average industry usability score of 68%. Only two participants result in a lower SUS score than the average performance. However, a participant shows a 100% usability score of the tested framework, and a general high interest can be observed among many participants in such a system.

TABLE 5. Comparison of performance with existing work.

EMI	ADA [59]	ECAs [16]
Accuracy performance of 69% in classifying depression symptom severity levels on the dataset, average sensitivity of 50%, specificity of 74%, the accuracy of 40% for experiments with real participants.	The first condition suggestion with severity was in accordance with the therapist's diagnosis in 44% of cases and 61% in the top five suggestions for real participants.	Average sensitivity (recall) of 50.3% for all three severity levels with mild symptoms being the lowest (21%), but specificity remained over 95% among patients.

TABLE 6. Comparison of usability with existing work.

EMI	ADA [59]	ECAs [16]
High usability 84.75% of the average SUS score, and good reliability as per Cronbach's alpha of 0.76.	High usability of 81.51% SUS score, however, face-to-face interviews were preferred over symptom checkers by 71% of participants.	Acceptability score 84.7%, good reliability according to Cronbach's alpha coef. 0.70.

Comparison analyses were performed among the existing literature that implemented chatbots in the mental health domain. Among the studies mentioned in the literature section, only a limited number of studies within this literature have employed chatbots to assess the severity levels of mental health issues, namely ADA [59] and Embodied Conversational Agents (ECAs) [16] and Babylon [15]. In this section, we undertake a comparative analysis of these studies. It is worth noting that Babylon Triage and Diagnostic System could not be included in our comparison due to its validation process, which involved assessing the virtual assistant's diagnostic performance against that of human doctors, while the experimental setup of our study was based on a dataset of real participants. Although this study [15] aims at detecting a wide range of medical conditions, experimental results for mental health-related problems were not provided, while this remains the main target of our work. Furthermore, the authors of the ECAs did not provide accurate results in the paper. Consequently, our comparison with ECAs is based on the evaluation metrics utilized in their statistical analysis, such as sensitivity and specificity.

As described in Table 5, ADA [59] correctly suggested the severity level of depression in 44% of the cases with real participants, while the top five suggestions improved the accuracy performance by 17%. Philip et al. [16] performed a stratified analysis of ECA to diagnose Major Depressive Disorder (MDD) among 35 participants compared to the diagnosis of the therapist. ECA showed a sensitivity of 21-73 among mild, moderate, and severe symptoms according to the severity function of BDI-II. Sensitivity improved with increasing severity of depressive symptoms, reaching 73% in people with severe depressive symptoms, but specificity remained greater than 95% for the three severity levels. Our EMI chatbot has achieved 40% to 65% of accuracy results when testing with real participants, while experiment results with the dataset (69%) strengthened our confidence in the

effectiveness of a dialogue system for the mental health domain.

Concerning usability performance, our experiment results are in line with previous related works. The usability of ADA was interpreted as high (mean 81.51, SD 11.82); however, more than two-thirds of the participants would have preferred face-to-face interviews with healthcare professionals, as shown in Table 6. On the other hand, the ECAs achieved high acceptability and reliability, and our chatbot demonstrates competitive performance (SUS score of 84.75%).

V. DISCUSSION

This work is unique in the related literature. Indeed, on the one hand, there have not been many publications performing such a comprehensive study of mental health problems in depth. On the other hand, although there are many applications considering depression detection, few of them focus on analyzing the level of mental health problems severity. By conducting experiments with real participants and testing our application, we can draw several observations. The chatbot framework cannot fully imitate and replace the medical professional, which was never the target of our work. However, we propose a different framework that, to the best of our knowledge, has not been proposed before, that is tailored for chatbots in mental health. Since the clinical interview scenario in the chatbot context still needs some improvements, it may not be sufficient to capture contextual information that clinical specialists may interpret during face-to-face interviews. For instance, the body language, facial expression, tone, and voice of the patient can provide substantial additional information from which the clinician can draw conclusions. A set of limitations apply to this work as well, as is explained later in this section.

Currently, we are in the process of Research Ethics Approval in collaboration with Danat Al Emarat Hospital for the recruitment of a clinical population. Therefore, we believe that our method could be used to detect signs of such conditions among individuals who are potentially at high risk of developing those mental health problems. This would allow us to explore and research other types of mental health disorders, such as generalized anxiety disorder, post-traumatic stress disorder (PTSD), etc. With the larger sample size from the hospital, we will have a chance of obtaining a high accuracy including multiple mental health conditions in addition to depression. Our approach may be useful in ensuring the mental well-being of the population at schools, educational institutes, and workplaces, where productivity is in the highest demand and is positively correlated with mental and physical characteristics.

From the clinical psychiatrist's viewpoint, the study holds promise as it explores the potential of utilizing chatbots in mental health services. One key advantage highlighted by the psychiatrist is the ability of chatbots to reduce the stigma associated with mental health problems. By offering online and easily accessible platforms for interactive assessment and screening, chatbots provide individuals with a discreet and convenient means to seek support. This can encourage more people to utilize the service, ultimately breaking down barriers and reaching a larger population. Moreover, the psychiatrist emphasizes the scalability of chatbots for mental health screening on a broader scale. By analyzing the collected data, it becomes possible to identify individuals who may require further clinical assessment and assistance. This data-driven approach facilitates early detection and intervention, ensuring appropriate steps are taken to support those with possible mental health problems.

The study potentially contributes to the field of mental health professionals, including clinical psychiatrists, by highlighting the advantages of incorporating such AI applications as chatbots into the mental health domain. The scalability of chatbot-based screening allows mental health professionals to reach a larger population and efficiently identify individuals who may require further clinical assessment and intervention. By analyzing the data generated from chatbot interactions, mental health professionals gain valuable insights into patterns, trends, and indicators of potential mental health issues. Moreover, the study's emphasis on chatbot technology highlights the potential for improved resource allocation and service planning within mental health settings. By gathering aggregated and anonymized data from chatbot interactions, mental health professionals can identify specific areas of need and tailor their services accordingly.

Through our planned future studies in collaboration with the hospital, which will grant us access to a larger and more diverse dataset, we anticipate an expanded array of opportunities to investigate a wide spectrum of mental health issues. This approach will help mitigate the limitations associated with our initially limited training and testing data.

While the potential benefits of utilizing AI in this field are vast, it is important to address the inherent privacy concerns that arise. As researchers delve into developing AI-powered applications, it becomes essential to conduct more enhanced testing before deploying these technologies in real-life settings. This includes an enhanced evaluation of data security, user consent, and the protection of sensitive personal information. Robust measures should be put in place to ensure that any data collected or processed by the AI system remains confidential. By conducting thorough testing and implementing privacy protocols, we can strive to strike the right balance between harnessing the power of AI in mental health and maintaining the privacy of users.

A. LIMITATIONS

During the development of this system, several potential limitations emerged that merit consideration for future enhancements. The chatbot's ability to detect emotions is restricted, and the absence of a comprehensive training dataset for Natural Language Understanding (NLU) impairs its conversational flexibility. Our experiments revealed a minor correlation between the length of user conversations and the accuracy of mental health condition classification.

Moreover, there is room for improvement in training chatbots to extract context from concise responses, and longer dialog exchanges could facilitate more intuitive user interactions. These observed gaps are predominantly attributed to the quality and quantity of the training dataset.

In addition, we acknowledge that the study sample size is relatively small. However, we have taken steps to address this limitation by planning to conduct a study in collaboration with a hospital where we will have access to a larger pool of patients under their care. Currently, we have submitted an ethics approval request to the hospital, and it is currently being processed.

The design choice of fewer categories used for our current study was made due to several reasons. First, by narrowing the scope to a smaller number of categories, we aimed to ensure the chatbot's ability to accurately detect these commonly occurring issues. A limited number of categories allows us to allocate more training data and resources to each category. Additionally, our study represents an initial exploration into the feasibility and effectiveness of using chatbots for mental health problem detection. Limiting the number of categories helped us assess the chatbot's performance in a controlled setting before potentially expanding to a wider range of mental health concerns and an increased number of categories.

We found that self-awareness is the key to solving issues. Many people lack understanding and awareness of their own mental health; therefore, they may not notice early signs of mental health issues, which may lead to more serious consequences. Due to their limited specificity, the PHQ-9 and PCL-C are better suited for initial screening than diagnostic evaluation. Therefore, the chatbot can be used for initial screening rather than diagnosis. We recognize the need for consistent assessment processes and datasets that enable a rigorous and fair comparison of various symptom checkers. Our experiments are in line with previous results in terms of the acceptance and potential benefits that users can obtain from AI-powered systems in mental health.

VI. CONCLUSION

In the last decade, mental health has garnered a considerable amount of academic and industry interest, demonstrating the high demand to investigate and implement systems with recent technological advancements. However, the accessibility and availability of such solutions are still matters for improvement. There is a high potential for AI systems to address these points with domain knowledge expertise.

In this research, our aim was to design a dialogue system for the detection of mental health problems that incorporates conversational AI to provide feedback on symptomatic predictions and recommendations, tailored to each individual's unique mental health profile, as determined through comprehensive analysis by the chatbot. The proposed system obtained high usability and good accuracy performance in our experimental study. This paper investigated the potential of artificial intelligence (AI) in conjunction with domain expertise to address the challenges associated with the detection of mental health problems. From a clinical psychiatrist's perspective, incorporating chatbots into mental health services holds promise for reducing stigma, improving accessibility, and streamlining the identification and support process for individuals with potential mental health concerns.

We have devised a methodology that allows us to capture information on the current mental health status of individuals. Moreover, considerable progress has been made in designing and developing the chatbot system that classifies the severity of two MHP and testing the framework's usability with real participants. Our contribution is to perform multiclass classification of the severity of mental illnesses, whereas the majority of the existing literature focuses on detecting the presence or absence of such problems. Consequently, we have obtained the dialogue system that provides feedback based on the chatbot's analysis. Additionally, our significant contribution to the field of mental health is the proposed digital twin framework, which holds great potential to provide innovative and personalized solutions to mental health challenges. It is also one of the few studies investigating performance evaluation on both existing datasets and real participants. The most important limitation lies in the lack of a clinically labeled dataset that would give more accurate performance results. Nonetheless, our findings are in support of the hypothesis that recent developments in natural language processing have promising results in assisting mental health professionals and increasing individuals' self-awareness about their mental health state. To the best of our knowledge, most chatbots fall short in providing crucial feedback and encouragement, a key feature that sets our chatbot apart and represents a promising advantage in delivering improved mental health outcomes.

REFERENCES

[1] *Depression and Other Common Mental Disorders: Global Health Estimates*, World Health Organization, Geneva, Switzerland, 2017.

[2] D. F. Santomauro et al., "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic," *Lancet*, vol. 398, no. 10312, pp. 1700–1712, 2021.

[3] T. Gaffney. (Oct. 2021). *Rates of Depression and Anxiety Climbed Across the Globe in 2020, Analysis Finds*. Accessed: Nov. 26, 2022. [Online]. Available: https://www.statnews.com/2021/10/08/mental-health-covid19-pandemic-global/

[4] V. Patel et al., "The lancet commission on global mental health and sustainable development," *Lancet*, vol. 392, no. 10157, pp. 1553–1598, 2018.

[5] M. Colizzi, A. Lasalvia, and M. Ruggeri, "Prevention and early intervention in youth mental health: Is it time for a multidisciplinary and trans-diagnostic model for care?" *Int. J. Mental Health Syst.*, vol. 14, no. 1, pp. 1–14, Dec. 2020.

[6] A. A. Abd-alrazaq, M. Alajlani, A. A. Alalwan, B. M. Bewick, P. Gardner, and M. Househ, "An overview of the features of chatbots in mental health: A scoping review," *Int. J. Med. Informat.*, vol. 132, Dec. 2019, Art. no. 103978.

[7] D. A. Regier, E. A. Kuhl, and D. J. Kupfer, "The DSM-5: Classification and criteria changes," *World Psychiatry*, vol. 12, no. 2, pp. 92–98, Jun. 2013.

[8] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial," *JMIR Mental Health*, vol. 4, no. 2, p. e19, Jun. 2017.

[9] S. Burchert, A. Kerber, J. Zimmermann, and C. Knaevelsrud, "Screening accuracy of a 14-day smartphone ambulatory assessment of depression symptoms and mood dynamics in a general population sample: Comparison with the PHQ-9 depression screening," *PLoS ONE*, vol. 16, no. 1, Jan. 2021, Art. no. e0244955.

[10] G. Cameron, D. Cameron, G. Megaw, R. Bond, M. Mulvenna, S. O'Neill, C. Armour, and M. McTear, "Towards a chatbot for digital counselling," in *Proc. Electron. Workshops Comput.*, 2017, pp. 1–7.

[11] A. Palanica, P. Flaschner, A. Thommandram, M. Li, and Y. Fossat, "Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey," *J. Med. Internet Res.*, vol. 21, no. 4, Apr. 2019, Art. no. e12887.

[12] A. Ahmed, N. Ali, S. Aziz, A. A. Abd-alrazaq, A. Hassan, M. Khalifa, B. Elhusein, M. Ahmed, M. A. S. Ahmed, and M. Househ, "A review of mobile chatbot apps for anxiety and depression and their self-care features," *Comput. Methods Programs Biomed. Update*, vol. 1, Jan. 2021, Art. no. 100012.

[13] T. W. Bickmore, S. E. Mitchell, B. W. Jack, M. K. Paasche-Orlow, L. M. Pfeifer, and J. O'Donnell, "Response to a relational agent by hospital patients with depressive symptoms," *Interacting Comput.*, vol. 22, no. 4, pp. 289–298, Jul. 2010.

[14] S. Hennemann, S. Kuhn, M. Witthöft, and S. M. Jungmann, "Diagnostic performance of an app-based symptom checker in mental disorders: Comparative study in psychotherapy outpatients," *JMIR Mental Health*, vol. 9, no. 1, Jan. 2022, Art. no. e32832.

[15] A. Baker, Y. Perov, K. Middleton, J. Baxter, D. Mullarkey, D. Sangar, M. Butt, A. DoRosario, and S. Johri, "A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis," *Frontiers Artif. Intell.*, vol. 3, Nov. 2020, Art. no. 543405.

[16] P. Philip, J.-A. Micoulaud-Franchi, P. Sagaspe, E. D. Sevin, J. Olive, S. Bioulac, and A. Sauteraud, "Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders," *Sci. Rep.*, vol. 7, no. 1, pp. 1–7, Feb. 2017.

[17] K. Denecke, S. Vaaheesan, and A. Arulnathan, "A mental health chatbot for regulating emotions (SERMO)—Concept and usability test," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 3, pp. 1170–1182, Jul. 2021.

[18] S. M. Jungmann, T. Klan, S. Kuhn, and F. Jungmann, "Accuracy of a chatbot (ADA) in the diagnosis of mental disorders: Comparative case study with lay and expert users," *JMIR Formative Res.*, vol. 3, no. 4, Oct. 2019, Art. no. e13863.

[19] D. Chambers, A. J. Cantrell, M. Johnson, L. Preston, S. K. Baxter, A. Booth, and J. Turner, "Digital and online symptom checkers and health assessment/triage services for urgent health problems: Systematic review," *BMJ Open*, vol. 9, no. 8, Aug. 2019, Art. no. e027743.

[20] H. L. Semigran, J. A. Linder, C. Gidengil, and A. Mehrotra, "Evaluation of symptom checkers for self diagnosis and triage: Audit study," *BMJ*, vol. 351, p. h3480, Jul. 2015.

[21] S. Gilbert, A. Mehl, A. Baluch, C. Cawley, J. Challiner, H. Fraser, E. Millen, M. Montazeri, J. Multmeier, F. Pick, C. Richter, E. Türk, S. Upadhyay, V. Virani, N. Vona, P. Wicks, and C. Novorol, "How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs," *BMJ Open*, vol. 10, no. 12, Dec. 2020, Art. no. e040269.

[22] A. Ceney, S. Tolond, A. Glowinski, B. Marks, S. Swift, and T. Palser, "Accuracy of online symptom checkers and the potential impact on service utilisation," *PLoS ONE*, vol. 16, no. 7, 2021, Art. no. e0254088.

[23] E. Bendig, B. Erb, L. Schulze-Thuesing, and H. Baumeister, "The next generation: Chatbots in clinical psychology and psychotherapy to foster mental health—A scoping review," *Verhaltenstherapie*, pp. 1–13, 2019.

[24] H. L. Semigran, D. M. Levine, S. Nundy, and A. Mehrotra, "Comparison of physician and computer diagnostic accuracy," *JAMA Internal Med.*, vol. 176, no. 12, p. 1860, Dec. 2016.

[25] G. Cameron, D. Cameron, G. Megaw, R. Bond, M. Mulvenna, S. O'Neill, C. Armour, and M. McTear, "Assessing the usability of a chatbot for mental health care," in *Proc. Int. Conf. Internet Sci.* Cham, Switzerland: Springer, 2018, pp. 121–132.

[26] J. Brooke, "SUS-A quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, no. 194, pp. 4–7, 1996.

[27] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," *J. Gen. Internal Med.*, vol. 16, no. 9, pp. 606–613, Sep. 2001.

[28] Gartner. *Digital Twin*. Accessed: Nov. 19, 2023. [Online]. Available: https://www.gartner.com/en/information-technology/glossary/digital-twin

[29] A. El Saddik, H. Badawi, R. A. M. Velazquez, F. Laamarti, R. G. Diaz, N. Bagaria, and J. S. Arteaga-Falconi, "DTwins: A digital twins ecosystem for health and well-being," *IEEE COMSOC MMTC Commun.*, vol. 14, no. 2, pp. 39–43, May 2019.

[30] A. El Saddik, "Digital twins: The convergence of multimedia technologies," *IEEE MultimediaMag.*, vol. 25, no. 2, pp. 87–92, Apr. 2018.

[31] H. F. Badawi, F. Laamarti, and A. El Saddik, "ISO/IEEE 11073 personal health device (X73-PHD) standards compliant systems: A systematic literature review," *IEEE Access*, vol. 7, pp. 3062–3073, 2019.

[32] F. Laamarti, H. F. Badawi, Y. Ding, F. Arafsha, B. Hafidh, and A. E. Saddik, "An ISO/IEEE 11073 standardized digital twin framework for health and well-being in smart cities," *IEEE Access*, vol. 8, pp. 105950–105961, 2020.

[33] B. Subramanian, J. Kim, M. Maray, and A. Paul, "Digital twin model: A real-time emotion recognition system for personalized healthcare," *IEEE Access*, vol. 10, pp. 81155–81165, 2022.

[34] M. N. Kamel Boulos and P. Zhang, "Digital twins: From personalised medicine to precision public health," *J. Personalized Med.*, vol. 11, no. 8, p. 745, Jul. 2021.

[35] J. D. Hamilton and L. Bickman, "A measurement feedback system (MFS) is necessary to improve mental health outcomes," *J. Amer. Acad. Child Adolescent Psychiatry*, vol. 47, no. 10, pp. 1114–1119, Oct. 2008.

[36] M. Spitzer, I. Dattner, and S. Zilcha-Mano, "Digital twins and the future of precision mental health," *Frontiers Psychiatry*, vol. 14, Mar. 2023, Art. no. 1082598.

[37] R. Ferdousi, F. Laamarti, M. A. Hossain, C. Yang, and A. El Saddik, "Digital twins for well-being: An overview," *Digit. Twin*, vol. 1, p. 7, Oct. 2021.

[38] E. Vildjiounaite, J. Kallio, J. Kantorovitch, A. Kinnula, S. Ferreira, M. A. Rodrigues, and N. Rocha, "Challenges of learning human digital twin: Case study of mental wellbeing: Using sensor data and machine learning to create HDT," in *Proc. 16th Int. Conf. Pervasive Technol. Rel. Assistive Environ.*, Jul. 2023, pp. 574–583.

[39] E. Miller, R. Hanlon, P. Lehrer, K. Mitchell, and M. Hancock, "Intelligent wellness," in *Proc. Int. Conf. Hum.-Comput. Interact.* Cham, Switzerland: Springer, 2023, pp. 232–249.

[40] B. Schrank, M. Amering, A. G. Hay, M. Weber, and I. Sibitz, "Insight, positive and negative symptoms, hope, depression and self-stigma: A comprehensive model of mutual influences in schizophrenia spectrum disorders," *Epidemiol. Psychiatric Sci.*, vol. 23, no. 3, pp. 271–279, Sep. 2014.

[41] K. Chakraborty and D. Basu, "Insight in schizophrenia-a comprehensive update," *German J. Psychiatry*, vol. 13, no. 1, pp. 17–30, 2010.

[42] Y. Windiatmoko, R. Rahmadi, and A. F. Hidayatullah, "Developing Facebook chatbot based on deep learning using RASA framework for university enquiries," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 1077, Feb. 2021, Art. no. 012060.

[43] BERT. *Huggingface Transformer Documentation*. Accessed: Jun. 27, 2023. [Online]. Available: https://huggingface.co/docs/transformers

[44] RASA. *Introduction to RASA Open Source & RASA Pro*. Accessed: Jun. 27, 2023. [Online]. Available: https://rasa.com/docs/rasa/

[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[46] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, and S. Marsella, "The distress analysis interview corpus of human and computer interviews," in *Proc. LREC*, Reykjavik, Iceland, 2014, pp. 3123–3128.

[47] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *J. Affect. Disorders*, vol. 114, nos. 1–3, pp. 163–173, Apr. 2009.

[48] Y. Wu et al., "Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: A systematic review and individual participant data meta-analysis," *Psychol. Med.*, vol. 50, no. 8, pp. 1368–1380, 2020.

[49] C. Shin, S.-H. Lee, K.-M. Han, H.-K. Yoon, and C. Han, "Comparison of the usefulness of the PHQ-8 and PHQ-9 for screening for major depressive disorder: Analysis of psychiatric outpatient data," *Psychiatry Invest.*, vol. 16, no. 4, pp. 300–305, Apr. 2019.

[50] M. Fadaee, A. Bisazza, and C. Monz, "Data augmentation for low-resource neural machine translation," 2017, *arXiv:1705.00440*.

[51] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," 2018, *arXiv:1805.06201*.
[52] J. Connell, A. O'Cathain, and J. Brazier, "Measuring quality of life in mental health: Are we asking the right questions?" *Social Sci. Med.*, vol. 120, pp. 12–20, Jan. 2014.
[53] H. Dalianis and H. Dalianis, "Evaluation metrics and evaluation," *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer, 2018, pp. 45–53.
[54] K. M. Ting, "Confusion matrix," in *Encyclopedia of Machine Learning and Data Mining*. ACM, 2010.
[55] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, no. 4857, pp. 1285–1293, Jun. 1988.
[56] M. Buckland and F. Gey, "The relationship between recall and precision," *J. Amer. Soc. Inf. Sci.*, vol. 45, no. 1, pp. 12–19, Jan. 1994.
[57] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, and N. Elhadad, "Diagnosis code assignment: Models and evaluation metrics," *J. Amer. Med. Inform. Assoc.*, vol. 21, no. 2, pp. 231–237, Mar. 2014.
[58] L. Derczynski, "Complementarity, F-score, and NLP evaluation," in *Proc. 10th Int. Conf. Lang. Resour. Eval.*, 2016, pp. 261–266.
[59] *Health. Powered by Ada*. Accessed: Jun. 27, 2023. [Online]. Available: https://ada.com/ -06-27.

AKBOBEK ABILKAIYRKYZY received the B.Sc. degree in automation and control from the Almaty University of Power Engineering and Telecommunications and the M.Sc. degree in machine learning from the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI). She is currently a Research Assistant with MBZUAI. With research interests in healthcare, energy, sustainability, digital transformation, and Metaverse domains, she brings a diverse perspective to her work. Prior to the master's degree, she was a Senior Automation and Control Engineer with a Copper Mining Corporation, Kazakhstan. Her achievements include first place in the Cisco Sustainability Challenge and a Finalist and a Silver Winner in the Gitex High Flyer x Ai Everything Hackathon.

FEDWA LAAMARTI received the Ph.D. degree in computer engineering from the University of Ottawa. She is currently a Research Fellow with the Mohamed bin Zayed University of Artificial Intelligence. Her research work was nominated for Best Thesis Award and she was a recipient of multiple prestigious scholarships, such as the Ontario Graduate Scholarship, the Queen Elizabeth Scholarship, and the University of Ottawa Excellence Scholarship. She also has several years of experience in the industry, in software design and analysis. She is working on multiple research projects with the Multimedia Communications Research Laboratory. Her research interests include the digital twin for health and well-being, artificial intelligence, and multimedia for social good.

MUFEED HAMDI received the bachelor's degree in medicine and surgery (MBChB) from Baghdad University, in 1980, the master's degree in psychiatry and neurology from the Faculty of Medicine, Ain-Shams University, Egypt, in 1990, and the Ph.D. degree in psychology from the College of Arts, Baghdad University, in 1998. He also completed a subspecialty training with the West Kent NHS and Social Care Trust, London, as an Honorary Specialist Registrar in learning disability psychiatry. Later in 2008, he has completed a six-month training course in "Global Mental Health: Trauma and Recovery Mastery Program" arranged by the Harvard Program in Refugee Trauma (HPRT) and Harvard Medical School. He has over 30 years of experience as a psychiatrist in both governmental and private sectors with reputable medical and health organizations in the United Arab Emirates and Iraq in addition to WHO.

ABDULMOTALEB EL SADDIK (Fellow, IEEE) was a Distinguished University Professor and the University Research Chair of the School of Electrical Engineering and Computer Science, University of Ottawa, before joining the Mohamed bin Zayed University of Artificial Intelligence. He was the Director of the Ottawa-Carleton Institute for Electrical and Computer Engineering (OCIECE), the Director of the Medical Devices Innovation Institute (MDII), and the Director of the Information Technology Cluster, Ontario Research Network on Electronic Commerce (ORNEC). He is a fellow of the Royal Society of Canada, a fellow of the Canadian Academy of Engineering, and a fellow of the Engineering Institute of Canada. He is an ACM Distinguished Scientist and he has received several awards, including the Friedrich Wilhelm Bessel Award from the German Humboldt Foundation and the IEEE Instrumentation and Measurement Society Technical Achievement Award.