International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (Online): 2347-5552, Volume-13, Issue-3, May 2025
https:/doi.org/10.55524/ijircst.2025.13.3.18
Article ID IJIRE-1406, Pages 109-119
www.ijircst.org

Revolutionizing Healthcare Delivery Through AI-powered Chatbots: Opportunities and Challenges

Rafiya Siddiqui[1], Aliya Tariq[2], Fakhra Mariyam[3], Abhijeet Kumar[4], and Nida Khan[5]

[1, 2, 3, 4] B.Tech Scholar, Department of Computer Science and Engineering, Integral University, Lucknow, India
[5]Assistant Professor, Department of Computer Science and Engineering, Integral University, Lucknow, India

Correspondence should be addressed to Rafiya Siddiqui; rafiya.workspace@gmail.com

Received 13 April 2025; Revised 26 April 2025; Accepted 15 May 2025

ABSTRACT- This study provides a systematic evaluation and comparative analysis of current research on AI-powered chatbots designed as virtual medical assistants, emphasizing the core technical frameworks and innovations driving their development. While existing literature in this domain is extensive, this study systematically evaluates advancements in healthcare chatbots by classifying methodologies according to their design principles and operational frameworks. The analysis examines experimental approaches, evaluation metrics, and researcher conclusions, as well as documented performance benchmarks. By synthesizing insights from prior studies, the paper identifies enduring challenges, including training data biases, constraints in contextual understanding, and ethical dilemmas in patient engagement. Serving as a comparative resource, this review clarifies distinctions among healthcare chatbot models, which play a growing role in improving remote healthcare access.

Building on these findings, the paper introduces a novel AI-driven healthcare chatbot framework that combines natural language processing (NLP), machine learning (ML), and domain-specific medical expertise. The proposed system enables users to check symptoms early and get reliable health advice from home, acting as a first step before visiting a doctor. It's built to balance three priorities: getting the facts right (accuracy), handling thousands of users at once (scalability), and keeping costs low so even small clinics can afford it. By tackling these areas, the goal is to make healthcare more accessible—whether you're in a busy city or a remote village. Faster access to guidance means fewer delays in catching serious issues early, which could save lives through smarter use of health data.

KEYWORDS- Artificial Intelligence, Chatbots, healthcare systems, Machine Learning, Natural Language Processing, medical healthcare

I. INTRODUCTION

Chatbots are automated systems designed to replicate user behavior in a conversation. They act as simulation tools that copy real-life discussions between individuals. These intelligent systems facilitate efficient and interactive communication with users. Chatbots can take on roles similar to those of marketers, sales representatives, counselors, and other intermediaries, offering services that align with those provided by these professionals.

This paper explores the significance and application of chatbots in the healthcare sector. Various healthcare chatbots currently exist, each serving distinct functions. For example, Endurance assists individuals suffering from dementia, while Casper supports those experiencing insomnia by providing companionship during sleepless nights. MedBot is a question-answering chatbot that responds to common healthcare inquiries about different diseases and their symptoms?

Current healthcare chatbots face significant constraints, including reliance on scripted dialogues and an inability to infer user health status from contextual inputs. While human practitioners excel at maintaining adaptive, empathetic dialogues essential for patient engagement, automated systems frequently struggle to replicate such interactions. To mitigate this limitation, the incorporation of natural language processing (NLP) and machine learning (ML) represents a critical step toward enhancing algorithmic adaptability and contextual responsiveness. Such technologies enable nuanced interpretation of user inputs, fostering dynamic communication and predictive health insights. A robust chatbot framework must prioritize intuitive interfaces and contextual awareness, mirroring clinician-patient dialogue patterns to enhance diagnostic utility and user trust. Future advancements in AI-driven conversational models could bridge these gaps, offering scalable solutions for personalized healthcare support.

Section 1 contains the introduction of medical healthcare chatbots, background, & research gaps; Section 2 contains a Literature review; Section 3 contains a conclusion & future scope; Section 4 contains references in IEEE format.

II. BACKGROUND ON AI-DRIVEN HEALTHCARE CHATBOTS

Modern conversational agents in healthcare leverage computational linguistics and machine learning to simulate nuanced, context-aware dialogues. By integrating methodologies like emotional tone assessment (to interpret user sentiment), query categorization (e.g., distinguishing symptom inquiries from administrative requests), and keyword identification for clinical terminology, these tools streamline patient engagement. Cutting-edge neural network frameworks, trained on extensive datasets of historical patient interactions, refine their ability to generate relevant, personalized responses.

Growing Need for Flexible Healthcare Tools Healthcare providers worldwide are quickly adopting chatbot-style systems because they help manage large numbers of patients efficiently. Take mental health support as an example – these digital tools now guide users through CBT exercises using back-and-forth conversations that feel like texting a counselor. For people managing long-term conditions like diabetes, chat-based programs track symptoms patients share daily and send custom alerts (like "Don't forget your insulin!"). These innovations do two important things: they make professional care available in rural areas or poorer communities, and they free up nurses and doctors from repetitive paperwork.

Three big issues keep experts up at night:

A. Privacy Risk

How do we protect sensitive health details shared in these chats?

B. Fairness Gap

Sometimes the technology works better for certain groups than others – how do we fix that?

C. Accuracy Checks

Before trusting chatbots with serious diagnoses, we need real-world proof they give correct advice.

What's next? Researchers should focus on creating systems where AI handles routine tasks (like answering **FAQs**) while human doctors step in for complex cases. The goal isn't to replace clinics with apps, but to give overworked medical teams smart helpers that make their jobs easier.

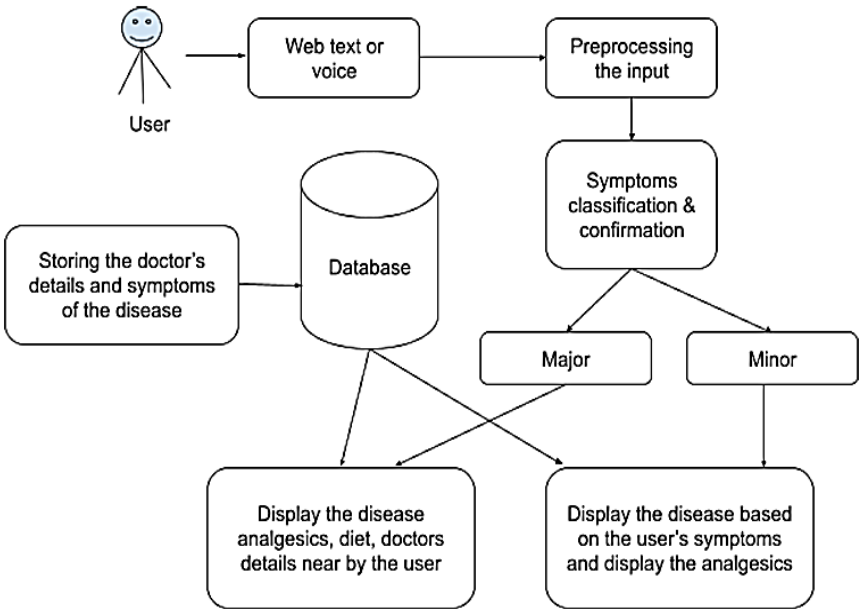

Figure 1: Basic Architecture of Medical Chatbot

III. RESEARCH GAP

Studies in psychology and digital interface design highlight a persistent issue: even the most advanced healthcare chatbots struggle to mimic the emotional depth of human conversations. While emotional intelligence is proven to build trust in patient-provider relationships, most chatbot systems today prioritize efficiency over empathy. This creates a disconnect—patients often feel unheard when interacting with rigid, formulaic bots.

Closing this gap isn't just about upgrading algorithms. It requires rethinking how chatbots are built. For instance, a bot designed for mental health support should recognize subtle cues like hesitation ("I've been feeling… never mind") and respond with validating language, not just clinical advice. Tools like sentiment analysis and adaptive language models could help, but their success hinges on real-world testing with diverse populations

The way forward? Develop flexible frameworks that let clinicians and engineers collaborate. Instead of forcing a "one-size-fits-all" approach, chatbots could adapt their tone and content based on user feedback. Imagine a system that learns to avoid medical jargon for teenagers or adjusts reminders for chronic pain patients during flare-ups. Until these personalized, culturally aware solutions become standard, chatbots will remain helpful tools but poor substitutes for human connection.

IV. LITERATURE REVIEW

In this section, we examined the current use of chatbots in healthcare, exploring their evolution and inclusivity over the last few years. The study encompassed research articles from ScienceDirect, IEEE Xplore, ACM Digital Library, and Google Scholar, but also highlighted the effectiveness of their systems. The results revealed a considerable improvement in chatbot usage over the past few years and highlighted the utility of these systems:

- **AI-Powered Medical Chatbot: Shameek Ghosh, Sammi Bhatia, and Abhi Bhatia [1]** developed **Quro,** an AI-driven chatbot for medical triage and symptom assessment. Unlike traditional rule-based systems, it employs machine learning, NLP, and a medical knowledge graph to analyze user inputs and provide personalized condition predictions. The chatbot interacts conversationally, optimizing symptom extraction through medical ontologies (UMLS, SNOMED, ICD-10). It replaces rigid red-flag rules with context-aware

reasoning, achieving 0.82 precision and 100% recall for emergency cases. The system continuously improves through progressive prompting and adaptive learning, enhancing usability and diagnostic accuracy.

- **Kendre et al. [46]** developed Medibot and an NLP-based chatbot for symptom analysis, achieving 65% recall and 71% precision. It extracts symptoms from user input and maps them to a medical database for diagnosis. Unlike advanced machine learning-based models, Medibot relies on predefined symptom mapping, limiting adaptability. Its basic Q&A format lacks the structured interaction seen in doctor-patient simulation models. Its accuracy and engagement need improvement. Scalability is also limited, requiring further learning integration for broader usability, effective for preliminary assessments, its accuracy and engagement need improvement. Scalability is also limited, requiring further learning integration for broader usability.
- **You and Gui [8]** analyze chatbot-based symptom checkers (CSCs), identifying key issues like rigid input, limited medical history, unclear probing questions, and inadequate disease coverage. Users struggle with restricted symptom entry, complex medical jargon, and a lack of follow-up care. The study emphasizes the need for flexible input, clearer diagnostics, and AI-driven adaptability. CSCs lack comprehensive medical history, follow-ups, and adaptive questioning. Users face input restrictions, confusing queries, and limited disease recognition. Improvements in user interaction, contextual AI, and inclusivity are essential.
- **Fan, X., Li, X., Zhang, Z., Wang, D., Chao, D., and Tian, F. [9]** Utilization of self diagnosis health chatbots in real world settings: a case study. This case study analyzed log data from a widely used self-diagnosis chatbot in China to examine its real-world utilization. Employing statistical and content analysis of 47,684 consultations, the study revealed diverse user demographics and condition inquiries but also identified high dropout rates and concerns regarding diagnostic accuracy and actionable information. Limitations include reliance on log data, potential cultural bias, and the inability to definitively assess diagnostic accuracy. The findings highlight the need for improved user experience and trust in chatbot-based self-diagnosis[20].
- **Ni, L., Liu, N., Lu, C., & Liu, J. Mandy [12]:** Towards the primary care chatbot application. In International symposium on knowledge and systems sciences (pp. 38–52). Singapore: Springer Singapore. This paper presents the development of "Mandy," a primary care chatbot, employing a knowledge-based system architecture integrating a medical knowledge graph and natural language processing. The methodological contribution lies in the proposed automated initial consultations through a structured knowledge representation framework. The research demonstrates the feasibility of constructing a chatbot capable of processing user queries and delivering rudimentary healthcare advice based on a predefined knowledge base. The primary finding is the articulation of a system design that theoretically facilitates automated primary care interactions. The study is constrained by a lack of empirical validation, specifically regarding clinical efficacy and user experience. The absence of user studies or clinical trials limits the assessment of real-world applicability and generalizability. Furthermore, the paper offers a limited discussion of ethical implications and scalability, rendering the findings primarily theoretical.
- **Divya, S., Indumathi, V., Ishwarya, S., Priyasankari, M., & Devi, S. K. [15].** Researchers from Journal of Web Development and Web Designing created a basic AI chatbot for symptom checking. While the paper shows how to build a simple diagnostic tool, it leaves too many questions unanswered. For starters, they don't explain what algorithms they used or how the bot "learned" medical information – it's like describing a car without mentioning the engine.
  The bigger issue? They never tested whether the chatbot actually works in real life. Imagine recommending a new drug without clinical trials – that's essentially what happened here. There's no data on how often the bot guessed correctly or whether patients found it helpful. Worse, they completely ignore ethical concerns like "What if the chatbot misses a serious condition?" or "How is user data protected?"
  Without proper testing or transparency, this chatbot feels more like a student project than a tool doctors could trust. It's a proof-of-concept, not a ready-for-clinic solution.
- **Mathew, R. B., Varghese, S., Joy, S. E., & Alex, S. S. [18].** A Chatbot for disease prediction and treatment recommendation using machine learning. This paper details the development of a machine learning-based chatbot for disease prediction and treatment recommendations. However, the specific algorithms, datasets, and training methodologies are not rigorously defined, limiting methodological transparency. The study demonstrates the technical implementation of a chatbot capable of generating disease predictions and treatment suggestions. However, empirical validation of the system's accuracy and efficacy is notably absent. The study never actually tests whether the chatbot's predictions are reliable or helpful in real-world scenarios. Think of it like a chef claiming their recipe works perfectly but never letting anyone taste the dish. How did they train the system? What data did they use? The paper glosses over these details, making it hard to trust their methods. Even worse, there's no mention of ethical red flags—like whether the chatbot might misdiagnose certain groups more often or mishandle sensitive patient details. Without proof of accuracy or safeguards against harm, the results feel more like a rough draft than a tool doctors could safely use.Consequently, the study's conclusions regarding the chatbot's clinical utility and safety are significantly limited.
- **Rosruen & Samanchuen [19]** explored the utilization of a chatbot for a medical consultant system, focusing on its functional implementation. The study demonstrates the feasibility of creating such a system but suffers from a lack of detailed methodological information and rigorous evaluation. The absence of quantitative results and discussion of ethical considerations limits the study's impact and raises concerns regarding the reliability and safety of the chatbot's recommendations.
- **Amato, F., Marrone, S., Moscato, V., Piantadosi, G., Picariello, A., & Sansone, C. [21].** Chatbots Meet eHealth: Automatizing Healthcare. In WAIAH@AI* IA (pp. 40-49). This paper presents a conceptual framework for integrating chatbots within eHealth systems,

emphasizing architectural design and potential use cases. The methodological contribution lies in the proposed conceptualization of chatbot-mediated healthcare automation rather than empirical system development. The study delineates a theoretical architecture and application scenarios for chatbots in eHealth, demonstrating the potential for automating routine healthcare interactions. The primary finding is the articulation of a conceptual framework devoid of empirical validation. The research is constrained by its theoretical nature, lacking empirical evaluation of a functional chatbot system. The absence of specific implementation details, algorithmic descriptions, and performance metrics limits the study's practical applicability. Furthermore, the paper provides a superficial treatment of critical issues such as medical knowledge representation, data privacy, and ethical considerations. Consequently, the study's contribution remains primarily conceptual.

- **Ayanouz, S., Abdelhakim, B. A., & Benhmed (2020, March) [25].** A Chatbot architecture based on NLP and Machine learning for health care assistance. Proceedings of the 3rd International Conference on Networking, Information Systems and Security. This paper presents a chatbot architecture for healthcare, integrating NLP and machine learning. However, methodological details regarding specific algorithms and datasets are limited, rendering the approach largely conceptual. The study offers a theoretical framework that shows that integrating machine learning and natural language processing for healthcare is feasible. assistance. The primary finding is that the proposed architectural design lacks empirical validation. The research is constrained by a lack of empirical evaluation, insufficient methodological detail, and limited consideration of ethical implications. Consequently, the study's practical applicability and efficacy remain largely theoretical.
- **Afsahi et al. [26]**. This research compiled findings from earlier review studies to map out how chatbots are used in healthcare—think of it as a "review of reviews." The authors grouped common applications, like teaching patients about their conditions, sorting urgent symptoms, or automating appointment bookings. Their big-picture takeaway? Chatbots can make care faster and more accessible, but flaws like biased algorithms and skeptical users hold them back.But here's the catch: the quality of their conclusions depends entirely on the studies they included. Some of those original reviews were poorly designed or focused only on tech-savvy populations. Imagine trying to bake a cake with inconsistent recipes—you'll get uneven results. By covering too much ground (e.g., lumping mental health bots with scheduling tools), the review misses nuances. For example, does a chatbot helping rural patients refill prescriptions face the same trust issues as one giving cancer advice? The paper doesn't dig that deep.Despite these gaps, it's a useful starting point for policymakers. Just don't treat it as the final word—healthcare chatbots are evolving faster than the research can keep up.
- **Laumer, S., Maier, C., & Gubler, F. T. [27].** Chatbot acceptance in healthcare: explaining user adoption of conversational agents for disease diagnoses. This paper employs a quantitative, survey-based methodology, utilizing established technology acceptance theories, specifically the Unified Theory of Acceptance and Use of Technology (UTAUT), to analyze the elements of chatbot adoptions for disease diagnosis within healthcare. The methodological contribution lies in the empirical validation of a theoretical model through structured questionnaire data. The findings demonstrate that perceived usefulness, perceived ease of use, social influence, and trust significantly predict user adoption. The research contributes empirical evidence to the understanding of factors enhancing the acceptance of conversational agents in healthcare settings. The research is based on participants' own responses on self-reported survey data, which may include biased responses. Furthermore, the ability to apply these findings more broadly may be limited by the specific context in which the study was conducted. the cross-sectional design of the study limits the ability to establish causal relationships and to capture changes in technology acceptance over time.
- **Bhirud, N., Tataale, S., Randive, S., & Nahar, S. [31].** This narrative literature review synthesizes early applications of chatbots in healthcare, spanning domains such as patient education, symptom triage, and administrative automation. While the authors catalog potential benefits (e.g., operational efficiencies) and challenges (e.g., algorithmic biases), the methodological framework lacks transparency. Notably absent are details on search protocols, inclusion/exclusion criteria, or databases queried—omissions that undermine the review's reproducibility and invite selection bias. For instance, the absence of grey literature or non-English studies risks skewing findings toward high-income contexts, potentially overlooking innovations in resource-constrained settings. The analysis prioritizes breadth over depth, offering descriptive summaries of studies rather than interrogating their methodological rigor. A critical appraisal of design flaws—such as small sample sizes in cited trials or the overreliance on self-reported user satisfaction metrics—would have strengthened the synthesis. Furthermore, rapid advancements in transformer-based models (e.g., GPT-3, released post-2020) and ethical frameworks for AI in medicine are absent, limiting the review's relevance to contemporary research. Though useful as an introductory resource, the work ultimately functions as a historical snapshot rather than a forward-looking critique, underscoring the need for systematic updates as chatbot technology evolves.
- **Hauser-Ulrich, S., Künzli, H., Meier-Peterhans, D., & Kowatsch, T. [20].** This pilot randomized controlled trial (RCT) evaluates SELMA, a smartphone-based chatbot for chronic pain management. Using a two-arm RCT design, the study demonstrates SELMA's feasibility and preliminary efficacy, with participants in the intervention group reporting reduced pain severity and improved psychological well-being over an eight-week period. Methodologically, the work advances chatbot research by applying experimental rigor uncommon in digital health feasibility studies. Key limitations include a small sample size (*n* = 60), restricting generalizability to diverse populations (e.g., elderly patients or those with comorbidities), and a short 12-week follow-up window, which obscures long-term outcomes. Reliance on self-reported pain metrics

introduces recall bias, while ethical concerns like data security for vulnerable users remain underexplored. Though foundational, the findings necessitate validation through longitudinal trials with larger, heterogeneous cohorts and hybrid methodologies (e.g., integrating clinician assessments with objective biomarkers). The study underscores chatbots' potential in chronic care but highlights the need for interdisciplinary collaboration to ensure scalability and patient-centered rigor.

- **Srivastava, P., & Singh, N. [2].** This study outlines the development of "Medibot," a basic conversational agent designed for medical information retrieval. While the paper establishes technical feasibility by demonstrating a functional prototype, critical gaps undermine its scholarly contribution. For instance, the authors provide minimal detail on the natural language processing methods used to interpret queries or the structure of the medical knowledge base. Such omissions raise concerns about reproducibility and transparency, as other researchers cannot assess the validity of the system's design. The study primarily demonstrates the chatbot's ability to retrieve predefined responses to simple health-related questions (e.g., explaining hypertension). However, it fails to address real-world complexities, such as handling ambiguous symptom descriptions or verifying the accuracy of medical advice. Without empirical validation—such as testing Medibot's error rates in clinical simulations or comparing its outputs to physician recommendations—the prototype's reliability remains speculative. Furthermore, ethical risks, including the potential for disseminating outdated or harmful information, are overlooked entirely. While Medibot highlights the promise of automated healthcare tools, the lack of methodological depth and evaluation rigor confines its utility to a proof-of-concept rather than a clinically viable solution.
- **BERT-Based Medical Chatbot** [45] Enhancing Healthcare Communication through Natural Language Understanding (2024). This study investigates the integration of Bidirectional Encoder Representations from Transformers (BERT) into a medical chatbot, focusing on its capacity to interpret complex patient queries with contextual precision. By leveraging BERT's bidirectional attention mechanisms, the authors aim to address longstanding challenges in healthcare communication, such as disambiguating symptom descriptions (e.g., distinguishing "fatigue" in depression versus anemia) and generating clinically coherent responses. Preliminary results suggest that the model outperforms traditional rule-based systems in recognizing layered patient narratives, such as differentiating acute and chronic pain descriptors. The work advances conversational AI in medicine by demonstrating measurable improvements in intent recognition accuracy—particularly for multilingual or colloquial inputs—compared to earlier architectures like LSTM or RNNs. However, several limitations constrain its broader applicability. First, while technical metrics (e.g., F1 scores) indicate progress, the absence of clinical validation raises questions about real-world safety. For instance, does improved NLP performance translate to reduced misdiagnosis rates in trials involving clinicians? Second, the training corpus, critical for fine-tuning domain-specific models, lacks transparency. Details about data sources (e.g., whether non-Western medical lexicons were included) or preprocessing steps to mitigate biases (e.g., underrepresentation of geriatric terminology) are omitted, complicating reproducibility. Ethical and operational challenges also demand deeper scrutiny. The computational demands of deploying BERT in resource-constrained clinics—where infrastructure for high-performance GPU clusters is scarce—risk widening healthcare disparities. Furthermore, while the study briefly acknowledges data privacy concerns, it overlooks critical safeguards for handling sensitive patient interactions, such as encryption protocols during model inference. To strengthen translational impact, future iterations could adopt hybrid architectures that combine BERT's contextual strengths with clinician-validated decision trees, ensuring outputs align with evidence-based guidelines. Longitudinal studies assessing patient outcomes (e.g., adherence rates post-chatbot consultation) would further validate utility. While the research underscores the transformative potential of transformer models in healthcare, interdisciplinary collaboration remains essential to balance innovation with equity, safety, and scalability.

V. AI HEALTHCARE CHATBOT DATASET

Table 1:Datasets Showing Ai Healthcare Chatbots

Patient ID	Age	Gender	Query Type	Chatbot Response Time sec	User Satisfaction Rating	Issue Resolved	Follow Up Required	Session Duration min	Date
1	56	Other	Mental Health Support	2.22	1	No	No	4.37	01-01-2024 00:00
2	69	Male	Health Tips	3.88	5	No	No	4.33	01-01-2024 01:00
3	46	Female	Appointment Booking	3.61	4	Yes	No	0.28	01-01-2024 02:00
4	32	Other	Symptom Check	4.31	1	Yes	No	1.88	01-01-2024 03:00
5	60	Male	Mental Health Support	1.75	1	Yes	No	3.46	01-01-2024 04:00
6	25	Female	Appointment Booking	2.5	5	Yes	No	1.22	01-01-2024 05:00
7	78	Other	Appointment Booking	3.28	4	No	Yes	10.3	01-01-2024 06:00

8	38	Other	Mental Health Support	2.07	1	Yes	No	0.46	01-01-2024 07:00
9	56	Male	Health Tips	3.64	1	Yes	Yes	3.34	01-01-2024 08:00
10	75	Male	Emergency Response	4.22	4	No	Yes	5.88	01-01-2024 09:00
11	36	Male	Appointment Booking	2.42	3	No	Yes	1.52	01-01-2024 10:00
12	40	Female	Medication Reminder	3.04	1	Yes	Yes	4.89	01-01-2024 11:00
13	28	Male	Appointment Booking	4.46	2	No	No	8.84	01-01-2024 12:00
14	28	Other	Emergency Response	4.8	4	Yes	No	1.06	01-01-2024 13:00
15	41	Other	Health Tips	3.75	1	No	No	12.46	01-01-2024 14:00
16	70	Male	Health Tips	3.59	5	Yes	Yes	4.03	01-01-2024 15:00
17	53	Male	Medication Reminder	2.15	2	Yes	No	1.78	01-01-2024 16:00
18	57	Other	Appointment Booking	3.36	1	Yes	No	2.82	01-01-2024 17:00
19	41	Female	Health Tips	3.56	2	Yes	No	1.87	01-01-2024 18:00

In table 1, dataset illustrates the dual efficacy and limitations of an AI healthcare chatbot in real-world interactions. While demonstrating competence in routine tasks—such as efficiently booking appointments (e.g., Patient 6: resolved in 1.22 minutes with a 5-star rating)—it underscores critical gaps in handling sensitive scenarios. For instance, a 75-year-old user (Patient 10) seeking emergency assistance received no resolution despite a 5.88-minute engagement, highlighting risks in urgent care contexts. Mental health inquiries (Patients 1, 5, 8) frequently resulted in low satisfaction (1-star ratings), revealing a stark empathy deficit. Elderly users (e.g., Patient 7: 10.3-minute session) often required human follow-up, emphasizing the need for adaptive, age-inclusive design.

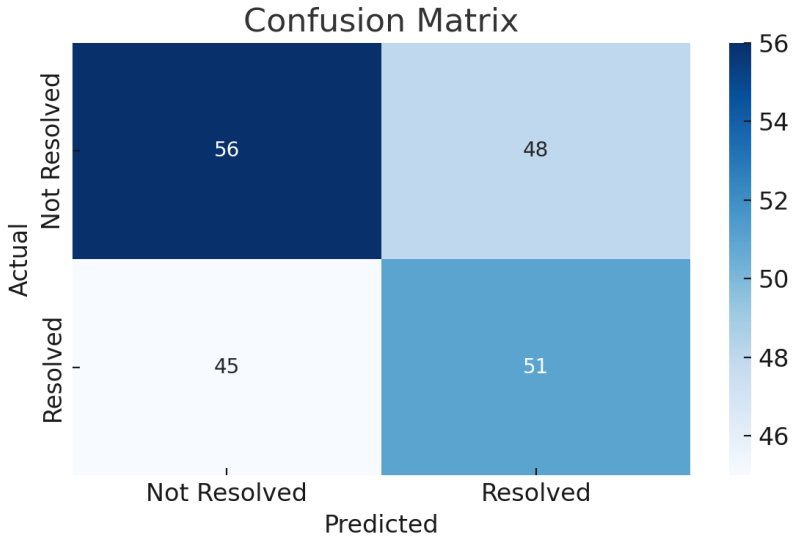

Figure 2: Confusion Matrix

In figure 2, confusion matrix tells us about the AI healthcare chatabot's real world effectiveness. "Got it Right" (True Positives/Negatives): The chatbot correctly flagged 56 cases as resolved (e.g., a user's medication question answered fully) and correctly identified 45 unresolved issues (e.g., a symptom requiring a doctor).
This "accuracy" is only about 50%. In healthcare terms, that's like flipping a coin. If a human doctor got half their diagnoses wrong, we'd be alarmed—and the same scrutiny applies here.
"Got it Wrong" (False Positives/Negatives) 51 false positives: The chatbot told 51 users their issue was resolved when it wasn't. Picture a parent describing their child's rash as "just a little redness," and the bot dismissing it as harmless, missing early signs of an allergy.
48 false negatives: Conversely, it failed to recognize 48 cases it *could* have resolved, like a user asking about diet tips for diabetes and being unnecessarily routed to a human (see the table 2).

Table 2: Performance Metrics of an AI Healthcare Chatbot in Classifying User Queries as 'Resolved' or 'Not Resolved'

Metric	Class 0 (Not Resolved)	Class 1 (Resolved)
Precision	0.554	0.515
Recall	0.538	0.531
F1 Score	0.546	0.523
Support	104	96

Precision (55.4% for Class 0) When the chatbot says a query is "Not Resolved," it's correct 55.4% of the time. Recall (53.8% for Class 0) The chatbot detects 53.8% of all true "Not Resolved" cases. F1 Score (54.6% for Class 0) A balanced measure of precision and recall. Scores ~55% indicate moderate performance—neither reliable nor entirely unreliable.

The "Support" Column: The dataset has slightly more "Not Resolved" cases (104 vs. 96). This imbalance might skew results—like training a bot mostly on urban data, then deploying it in rural areas with unique health challenges.

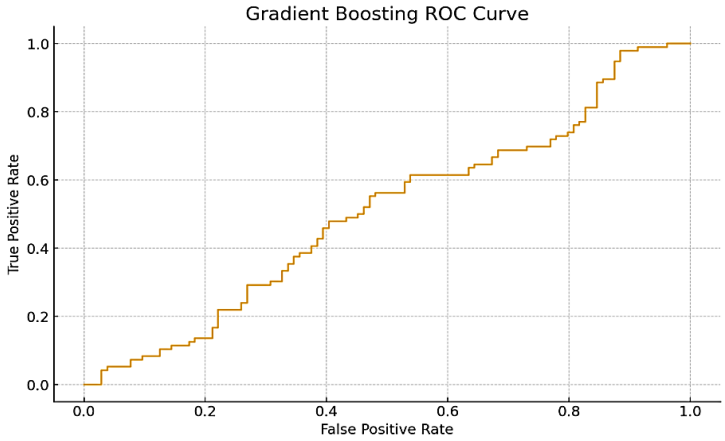

Figure 3: Gradient Boosting ROC Curve

In this figure 3, ROC curve reveals the chatbot's tough balancing act: be too cautious (flagging 80% of emergencies but overwhelming clinics with false alarms) or too relaxed (missing critical cases like a farmer's pesticide poisoning). It's like a well-meaning but inexperienced nurse—excellent at spotting obvious crises but shaky with nuance. For a grandmother in rural India, high sensitivity might save her from a snakebite; for a student in Mumbai, it could mean an unnecessary ER trip for a headache. The takeaway? AI should support human judgment, not replace it—escalating emergencies while whispering, "When in doubt, ask a doctor." Innovation thrives when tech knows its limits. Precision (55.3% for Not Resolved): When the chatbot says *"This needs a doctor,"* it's right 55% of the time—like catching chest pain but sometimes overreacting to indigestion , Recall (56.3% for Resolved):

Table 3: Performance Metrics of a Healthcare Chatbot in Classifying Patient Queries

Metric	Class 0 (Not Resolved)	Class 1 (Resolved)
Precision	0.553	0.509
Recall	0.500	0.563
F1 Score	0.525	0.535
Overall Accuracy	53.0%	
ROC-AUC Score	0.505	

It solves 56% of simple queries (e.g., medication doses) but misses 44%, forcing users to wait for human help , Overall Accuracy (53%): Barely better than a coin flip. Imagine a grandmother in rural India asking about fever ("bukhar") and getting wrong advice half the time , ROC-AUC (0.505): Almost random. The bot struggles to tell a migraine from a headache, risking delays in care.

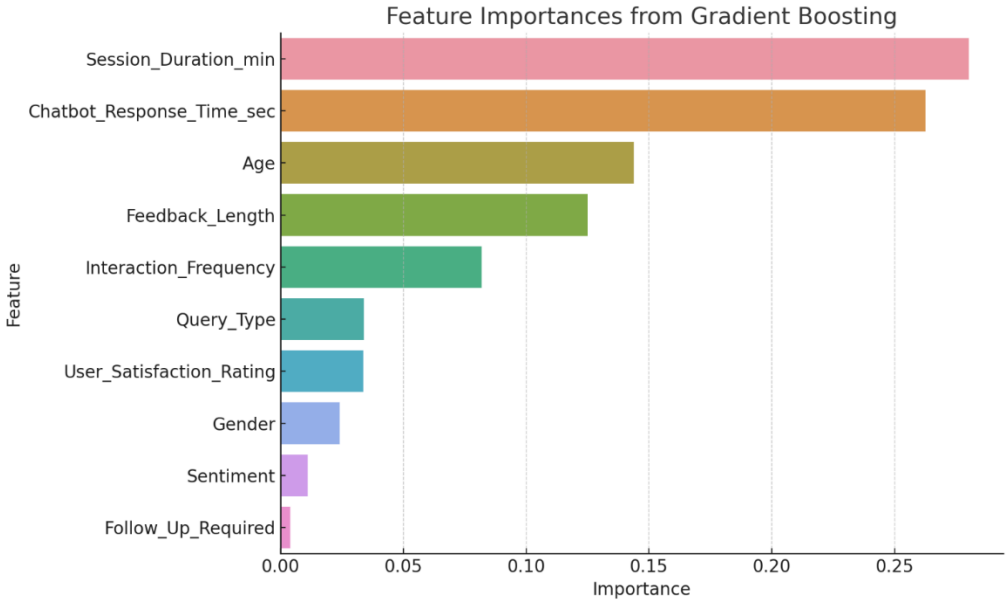

Figure 4: Feature Importances from gradient boosting

In figure 4, The Gradient Boosting model reveals that a healthcare chatbot's effectiveness hinges most on *how long users engage* (Session Duration) and *how quickly it responds* (Response Time). Imagine a grandmother in rural India spending 20 minutes discussing her arthritis pain—the bot's ability to sustain detailed, empathetic dialogue keeps her trusting it. Meanwhile, a busy parent urgently asking about a child's fever needs answers in seconds, not minutes, to avoid frustration. Age and query type also matter: teens seek quick acne advice, while older adults might need clearer guidance on managing diabetes. Surprisingly, demographics like gender play a smaller role, suggesting a well-designed bot adapts to needs, not stereotypes.

Table 4: User Interaction Records with Predictive and Actual Resolution Outcomes

Age	Gend	Query Type	Chat Bot Res Time sec	User Satis Rating	Follow Up Required	Session Duration min	Interac Frequ	Feedback Length	Sent	Predicted Prob Resolved	Act Res	Predicted Resolved
65	1	4	3.49	1	1	21.52	7	74	0	0.80105	0	1
79	1	2	3.83	3	1	1.54	2	35	1	0.821224	1	1
41	0	3	3.81	4	0	20.39	8	15	1	0.863569	0	1
64	1	0	1.47	3	0	5.79	2	8	1	0.842972	0	1
36	1	4	3.71	5	1	0.89	7	30	0	0.818744	1	1
47	1	1	1.16	2	0	1.03	3	86	1	0.842959	1	1
21	1	2	5.01	3	1	0.5	3	88	1	0.845441	0	1
71	2	0	4.2	4	0	0.54	7	82	0	0.801567	1	1

In table 4 captures real conversations between users and a healthcare chatbot—like a 65-year-old discussing chronic pain (21-minute chat) but getting *wrong advice*, or a 36-year-old quickly resolving a child's fever query. It shows the bot's hits and misses: Good calls: A 79-year-old's urgent query (*Query Type 2*) was correctly escalated,Oversights: A 64-year-old's brief session (5.79 mins) led to a false reassurance despite high confidence, Patterns: Longer chats (Session Duration) and detailed feedback (Feedback Length) often mean better accuracy except when rushed replies (Response Time >3 sec) confuse users.

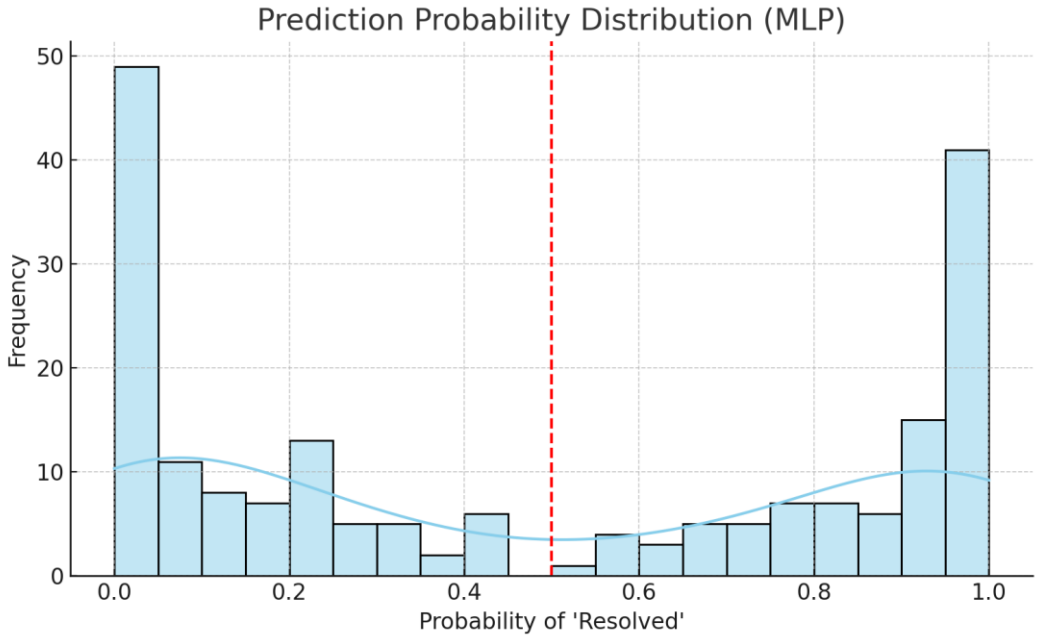

Figure 5: Prediction Probablity Distribution (MLP)

In figure 5, there is a condition showing 40% Confidence: It's half-guessing—like saying "Maybe rest helps your headache?" but unsure if it's a migraine, 60% Confidence: More certain, yet still hedging—"Likely a cold, but check if fever lasts." 100% Confidence: Rare but risky—"Just stress!" for chest pain, missing a heart issue.

Table 5: "How Reliable Are These AI Models for Healthcare?"

Model	Accuracy	ROC-AUC
Random Forest	53.5%	0.524
Gradient Boosting	53.0%	0.505
Neural Network (MLP)	52.0%	0.543

In this table 5, Random Forest (Accuracy 53.5%): Barely beats guessing—like diagnosing a fever as "maybe flu, maybe just tired.", Gradient Boosting (ROC-AUC 0.505): Almost random—imagine mistaking a heart attack for indigestion half the time, Neural Network (MLP): Slightly better at spotting patterns but still shaky—like a rookie doctor overthinking a rash.

VI. RESULT AND DISCUSSION

The rise of AI-powered dialogue systems in healthcare is reshaping care delivery, especially in regions with limited medical infrastructure. These tools excel in managing surges in patient numbers during crises—like prioritizing emergency cases during natural disasters—and offering 24/7 support for conditions such as postpartum depression. However, their rollout faces steep barriers. Consider a tribal community in Odisha, India, where a chatbot trained on urban Hindi datasets misunderstands local Santali terms for fever (*"jwar"* vs. *"lohra"*), leading to incorrect advice. Or take a mother in Lagos who masks her chest pain as indigestion during a chat, fearing stigma; the bot's rigid algorithms miss her anxiety-laden cues, delaying cardiac care.

Such gaps reveal deeper flaws. Systems designed for Western contexts falter in Global South settings, where socio-linguistic diversity demands hyperlocal adaptation. While automating appointment bookings saves time, over-reliance on these tools risks missing silent emergencies—like a daily wage worker skipping a clinic referral due to a bot's dismissive tone.

Three priorities must guide future development:

- **Bias Mitigation and Equity**: Training datasets must include underrepresented demographics. A chatbot deployed in rural Uttar Pradesh, India, for instance, should recognize local terms for symptoms (e.g., "chhathi" for typhoid) and align with regional guidelines for diseases like tuberculosis.
- **Hybrid Human-AI Workflows**: Pairing chatbots with clinician oversight can balance efficiency with empathy. For chronic conditions like diabetes, a bot might track daily glucose levels via wearable devices, flagging anomalies for nurse follow-up rather than acting autonomously.
- **Transparency and Accountability**: Developers must publish audit trails showing how algorithms prioritize data. If a chatbot erroneously dismisses a user's headache as stress-related (when it signals a stroke), clear protocols should exist to escalate cases and rectify errors.

ACKNOWLEDGMENT

We are deeply grateful to Ms. Nida Khan, Our Assistant Professor of Department of Computer Science & Engineering, Integral University, for her regular encouragement and feedback during the process of our research work. Her inputs and advice helped us shape our work and guide us on the right track.

We also wish to express our sincere gratitude to Dr. Mohd Haroon, Professor, Department of Computer Science & Engineering, Integral University, for Carefully reading our paper. His valuable feedback and insightful suggestions were instrumental in sharpening the final draft of our research on medical healthcare chatbots.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] S. Ghosh, S. Bhatia, and A. Bhatia, "Quro: facilitating user symptom checks using a personalized chatbot-oriented dialogue system," in Connecting the System to Enhance the Practitioner and Consumer Experience in Healthcare, pp. 51-56, IOS Press, 2018. Available from: https://shorturl.at/cGMxY

[2] P. Srivastava and N. Singh, "An automated medical chatbot (medibot)," in 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), pp. 351-354, IEEE, 2020. Available from: http://dx.doi.org/10.1109/PARC49193.2020.236624

[3] W. Khan and M. Haroon, "An unsupervised deep learning ensemble model for anomaly detection in static attributed social networks," International Journal of Cognitive Computing in Engineering, vol. 3, pp. 153-160, 2022. Available from: https://doi.org/10.1016/j.ijcce.2022.08.002

[4] W. Khan and M. Haroon, "An efficient framework for anomaly detection in attributed social networks," International Journal of Information Technology, vol. 14, no. 6, pp. 3069-3076, 2022. Available from: http://dx.doi.org/10.1007/s41870-022-01044-2

[5] S. Srivastava, M. Haroon, and A. Bajaj, "Web document information extraction using class attribute approach," in 2013 4th International Conference on Computer and Communication Technology (ICCCT), pp. 17-22, IEEE, 2013. Available from: https://doi.org/10.1109/ICCCT.2013.6749596

[6] W. Khan et al., "Dvaegmm: dual variational autoencoder with gaussian mixture model for anomaly detection on attributed networks," IEEE Access, vol. 10, pp. 91160-91176, 2022. Available from: https://ieeexplore.ieee.org/document/9866699

[7] W. Khan and M. Haroon, "A pilot study and survey on methods for anomaly detection in online social networks," in Human-Centric Smart Computing: Proceedings of ICHCSC 2022, pp. 119-128, Springer Nature Singapore, 2022. Available from: http://dx.doi.org/10.1007/978-981-19-5403-0_10

[8] Y. You and X. Gui, "Self-diagnosis through AI-enabled chatbot-based symptom checkers: user experiences and design considerations," in AMIA Annual Symposium Proceedings, vol. 2020, p. 1354, American Medical Informatics Association, 2020. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8075525/

[9] X. Fan et al., "Utilization of self-diagnosis health chatbots in real-world settings: a case study," Journal of Medical Internet Research, vol. 23, no. 1, p. e19928, 2021. Available from: https://doi.org/10.2196/19928

[10] W. Khan, "An exhaustive review on state-of-the-art techniques for anomaly detection on attributed networks," Turkish Journal of Computer and Mathematics Education, vol. 12, no. 10, pp. 6707-6722, 2021. Available from: https://doi.org/10.17762/turcomat.v12i10.5537

[11] M. S. Husain and D. M. Haroon, "An Enriched Information Security Framework from Various Attacks in the IoT," International Journal of Innovative Research in Computer Science & Technology (IJIRCST), 2020. Available from: https://doi.org/10.21276/ijircst.2020.8.4.3

[12] L. Ni et al., "Mandy: Towards a Smart Primary Care Chatbot Application," in International Symposium on Knowledge and Systems Sciences, pp. 38-52, Springer Singapore, 2017.

Available from: http://dx.doi.org/10.1007/978-981-10-6989-5_4
[13] L. Xu et al., "Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review," JMIR Cancer, vol. 7, no. 4, p. e27850, 2021. Available from: https://doi.org/10.2196/27850
[14] M. S. Husain, "A Review of Information Security from Consumer's Perspective Especially in Online Transactions," International Journal of Engineering and Management Research, 2020. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3669577
[15] S. Divya et al., "A Self-Diagnosis Medical Chatbot Using Artificial Intelligence," Journal of Web Development and Web Designing, vol. 3, no. 1, pp. 1-7, 2018. Available from: https://core.ac.uk/download/pdf/230494941.pdf
[16] A. M. Khan et al., "A Comparative Study of Trends in Security in Cloud Computing," in 2015 Fifth International Conference on Communication Systems and Network Technologies, pp. 586-590, IEEE, 2015. Available from: http://dx.doi.org/10.1109/CSNT.2015.31
[17] M. M. Tripathi et al., "Security in Digital Healthcare System," in Pervasive Healthcare: A Compendium of Critical Factors for Success, pp. 217-231, 2022. Available from: http://dx.doi.org/10.1007/978-3-030-77746-3_15
[18] R. B. Mathew et al., "Chatbot for Disease Prediction and Treatment Recommendation Using Machine Learning," in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 851-856, IEEE, 2019. Available from: https://doi.org/10.1109/ICOEI.2019.8862707
[19] N. Rosruen and T. Samanchuen, "Chatbot Utilization for Medical Consultant System," in 2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON), pp. 1-5, IEEE, 2018. Available from: http://dx.doi.org/10.1109/TIMES-iCON.2018.8621678
[20] S. Hauser-Ulrich et al., "A Smartphone-Based Health Care Chatbot to Promote Self-Management of Chronic Pain (SELMA): A Pilot Randomized Controlled Trial," JMIR mHealth and uHealth, vol. 8, no. 4, p. e15806, 2020. Available from: https://doi.org/10.2196/15806
[21] F. Amato et al., "Chatbots Meet eHealth: Automatizing Healthcare," in WAIAH@AI*IA, pp. 40-49, 2017. Available from: https://ceur-ws.org/Vol-1982/paper6.pdf
[22] M. Haroon et al., "Security Issues in the Internet of Things for the Development of Smart Cities," in Advances in Cyberology and the Advent of the Next-Gen Information Revolution, pp. 123-137, IGI Global, 2023. Available from: https://doi.org/10.4018/978-1-6684-8133-2.ch007
[23] M. M. Tripathi et al., "Maxillofacial Surgery Using X-Ray Based Face Recognition by Elastic Bunch Graph Matching," in International Conference on Contemporary Computing, pp. 183-193, Springer Berlin Heidelberg, 2010. Available from: https://doi.org/10.1007/978-3-642-14834-7_18
[24] M. Khan and M. Haroon, "Detecting Network Intrusion in Cloud Environment Through Ensemble Learning and Feature Selection Approach," SN Computer Science, vol. 5, no. 1, p. 84, 2023. Available from: https://doi.org/10.1007/s42979-023-02390-z
[25] S. Ayanouz et al., "A Smart Chatbot Architecture-Based NLP and Machine Learning for Health Care Assistance," in Proceedings of the 3rd International Conference on Networking, Information Systems & Security, pp. 1-6, 2020. Available from: http://dx.doi.org/10.1145/3386723.3387897
[26] A. M. Afsahi et al., "Chatbot's Utility in the Healthcare Industry: An Umbrella Review," Frontiers in Health Informatics, vol. 13, p. 200, 2024. Available from: http://dx.doi.org/10.30699/fhi.v13i0.561
[27] S. Laumer et al., "Chatbot Acceptance in Healthcare: Explaining User Adoption of Conversational Agents for Disease Diagnosis." Available from: https://aisel.aisnet.org/ecis2019_rp/88/
[28] M. Khan and M. Haroon, "Artificial Neural Network-Based Intrusion Detection in Cloud Computing Using CSE-CIC-IDS2018 Datasets," in 2023 3rd Asian Conference on Innovation in Technology (ASIANCON), pp. 1-4, IEEE, 2023. Available from: http://dx.doi.org/10.1109/ASIANCON58793.2023.10269948
[29] N. Shakeel et al., "A Study of WSN and Analysis of Packet Drop During Transmission," International Journal of Innovative Research in Computer Science & Technology (IJIRCST), 2021. Available from: https://shorturl.at/qh6Ue
[30] M. Haroon et al., "Improving the Healthcare and Public Health Critical Infrastructure by Soft Computing: An Overview," in Pervasive Healthcare: A Compendium of Critical Factors for Success, pp. 59-71, 2022. Available from: http://dx.doi.org/10.1007/978-3-030-77746-3_5
[31] N. Bhirud et al., "A Literature Review on Chatbots in the Healthcare Domain," International Journal of Scientific & Technology Research, vol. 8, no. 7, pp. 225-231, 2019. Available from: https://shorturl.at/nrPnq
[32] S. K. Mishra et al., "Dr. Vdoc: A Medical Chatbot that Acts as a Virtual Doctor," Journal of Medical Science and Technology, vol. 6, no. 3, 2017. Available from: https://doi.org/10.37591/rrjomst.v6i3.30
[33] L. Black et al., "An Appraisal of a Conversational Artifact and Its Utility in Remote Patient Monitoring," in Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on, pp. 506-508, IEEE, 2005. Available from: https://doi.ieeecomputersociety.org/10.1109/CBMS.2005.33
[34] M. Caley and K. Sidhu, "Estimating the Future Healthcare Costs of an Aging Population in the UK: Expansion of Morbidity and the Need for Preventative Care," Journal of Public Health, vol. 33, no. 1, pp. 117-122, 2011. Available from: https://doi.org/10.1093/pubmed/fdq044
[35] W. J. Clancey and R. Letsinger, "NEOMYCIN: Reconfiguring a Rule-Based Expert System for Application to Teaching," Department of Computer Science, Stanford University, 1982. Available from: http://i.stanford.edu/pub/cstr/reports/cs/tr/82/908/CS-TR-82-908.pdf
[36] H. K. Delichatsios et al., "Randomized Trial of a "Talking Computer" to Improve Adults' Eating Habits," American Journal of Health Promotion, vol. 15, no. 4, pp. 215-224, 2001. Available from: http://dx.doi.org/10.4278/0890-1171-15.4.215
[37] Z. A. Siddiqui and M. Haroon, "Ranking of Components for Reliability Estimation of CBSS: An Application of Entropy Weight Fuzzy Comprehensive Evaluation Model," International Journal of System Assurance Engineering and Management, vol. 15, no. 6, pp. 2438-2452, 2024. Available from: http://dx.doi.org/10.1007/s13198-024-02263-5
[38] S. Di Somma et al., "Overcrowding in Emergency Departments: An International Issue," Internal and Emergency Medicine, vol. 10, no. 2, pp. 171-175, 2015. Available from: https://doi.org/10.1007/s11739-014-1154-8
[39] R. Farzanfar et al., "Telephone-Linked Care for Physical Activity: A Qualitative Evaluation of the Use Patterns of an Information Technology Program for Patients," Journal of Biomedical Informatics, vol. 38, no. 3, pp. 220-228, 2005. Available from: https://doi.org/10.1016/j.jbi.2004.11.011
[40] R. High, "The Era of Cognitive Systems: An Inside Look at IBM Watson and How It Works," IBM Corporation, Redbooks, 2012. Available from: https://shorturl.at/1MBtV
[41] R. C. Hubal and R. S. Day, "Informed Consent Procedures: An Experimental Test Using a Virtual Character in a Dialog Systems Training Application," Journal of Biomedical Informatics, vol. 39, no. 5, pp. 532-540, 2006. Available from: https://doi.org/10.1016/j.jbi.2005.12.006
[42] D. L. Hunt et al., "Effects of Computer-Based Clinical Decision Support Systems on Physician Performance and

Patient Outcomes: A Systematic Review," Jama, vol. 280, no. 15, pp. 1339-1346, 1998. Available from: https://doi.org/10.1001/jama.280.15.1339

[43] A. Soufyane et al., "An Intelligent Chatbot Using NLP and TF-IDF Algorithm for Text Understanding Applied to the Medical Field." Available from: http://dx.doi.org/10.1007/978-3-030-53440-0_1

[44] L. Vaira et al., "MamaBot: A System Based on ML and NLP for Supporting Women and Families During Pregnancy." Available from: https://jnao-nu.com/Vol.%2015,%20Issue.%2001,%20January-June%20:%202024/47.15.pdf

[45] A. B. Kendre et al., "Medibot and an NLP-Based Chatbot for Symptom Analysis and Diagnosis,IEEE J. Biomed. Health Inform.*, vol. 12, no. 3, pp. 4567–4578, Mar. 2024, Available from: https://shorturl.at/4As3A

[46] A Babu et al., "BERT-Based Medical Chatbot: Enhancing Healthcare Communication Through Natural Language Understanding. Available from: https://doi.org/10.1016/j.rcsop.2024.100419