A Structured and Safety -Aware AI Framework for Digital Health Consultation

Arulprasanth S[1]*, Amudhan P[1], Sherly Puspha Annabel L[1]
[1] *Department of Artificial Intelligence and Data Science, St. Joseph's College of Engineering, Chennai, India*
s.arulprasanth2004@gmail.com, adsamudhan@gmail.com, sherlyannabel@gmail.com
*Corresponding Author: s.arulprasanth2004@gmail.com

Abstract. The increasing use of digital health consultation systems shows that technology now plays an essential role in identifying health problems and teaching patients. The existing digital health systems depend entirely on text communication which creates problems for users who need assistance and for health information that needs to be shared. Actual systems lack the ability to show users the visual aspects of their symptoms along with their emotional experiences which results in decreased system performance. This research presents an AI-based digital health consultation system which uses real-time avatar interaction together with organized symptom assessment to help users complete their medical consultation. The framework uses clinical reasoning principles to enable users to submit symptoms through text and audio and image methods. Virtual avatars with synchronized lip movements and changing facial expressions help users understand information better while increasing their interaction with the system. The proposed system was evaluated using a disease categorization module that classifies symptoms across five health categories, achieving a mean classification accuracy of 94.9% with low response latency. The framework provides accurate health assessment which uses evidence-based information to create non-diagnostic health assessment results according to experimental test results and confusion matrix evaluation.

Keywords: Digital Health Consultation, Structured Symptom Analysis, Multimodal Health Data, Safety-Aware AI, Avatar-Based Interaction, Preliminary Medical Guidance, Responsible Healthcare AI, Clinical Decision Support

1 INTRODUCTION

The increasing use of digital health consultation systems has established new pathways to enhance healthcare accessibility while promoting early detection of medical symptoms [1]. The systems permit users to explain their medical problems which results in them receiving initial medical information without needing direct physician contact [2]. Most current conversational health tools provide users with unstructured text-based interactions but they do not support a structured medical consultation process which uses clinical reasoning [3].

This has significant consequences. Users will miss vital information regarding their symptoms since they will have no official steps to follow when searching for answers to their questions [4]. Existing systems will not have visual or emotional dimensions; therefore, they will not communicate well since they cannot show the users any visual symptoms to aid in diagnosis [5].

R. Vasanth Kumar Mehta et al. (eds.), *Proceedings of the International Conference on Intelligent Systems for a Sustainable Future (ISSF 2026)*, Atlantis Highlights in Intelligent Systems 16,
https://doi.org/10.2991/978-94-6239-693-7_13

This is because the research shows that the use of multimodal input that combines text and speech and images with virtual agents that have expressiveness capabilities will improve user understanding and retention in digital health environments [6]. The current platforms use separate systems to run their different parts of operation as opposed to creating a single system that integrates all their parts and they lack effective safety measures that can prevent users from providing overconfident or deceptive answers [7].

This paper proposes a structured and safety-focused digital health consultation framework to bridge the gaps mentioned above. System Overview: The proposed framework includes a multi-modal symptom acquisition tool, a structured clarification tool, a contextual symptom interpretation engine, and a risk-aware guidance generation tool, which are delivered to the user through a lip-synced avatar with adaptive emotions. The proposed system aims to achieve the goal of accurate health prediction and understanding at an early stage with appropriate ethical and safety guidelines.

2 RELATED WORK

Traditional health consultation systems were rule-based systems that focused on specific standard questionnaires with strict decision trees [1]. They offered definite results but required users to describe their symptoms using a limited set of words and had difficulties with diseases with similar symptom profiles [3].

NLP and machine learning introduced the conversational health assistants that would interpret the user input [4]. This technology had its challenges as well because most of them didn't run with a full session-level model but only interpreted the latest user input while ignoring the context of the symptom and provided contradictory information [7]. The absence of safety filtering on the assistants made people worried that users would place too much trust in auto-medical responses [9].

Recent studies on the application of large language models within the healthcare domain have revealed promising levels of conversational ability and context sensitivity [13]. However, these studies have also shown that, unless constrained, the models are able to provide confident assertions on output that are clinically incorrect [2]. Therefore, the application of these models must be designed with specific workflows to impose responsibility on the model through the limitation of the generated output, rather than the diagnosis itself [1].

Research carried out on virtual avatars in the context of virtual consultation environments has suggested that the use of expressive avatars, with facial expressions and dynamic synchronization with mood (DD) and emotion, can contribute to the comfort of users and the clarity of the consultation process [14][15]. However, the existing systems seem to view avatars as separate interface components, rather than as emotionally rational agents that are part of the consultation logic, with the capacity to provide only a limited number of predefined emotions [16]. The proposed framework addresses both of these by embedding the avatar-mediated delivery within a single, safety-governed consultation pipeline.

3 METHODOLOGY

This framework allows for safe, structured, multimodal digital health consultations with patients. By enforcing a turn-taking interaction sequence, rather than tolerating anything like free-form conversation, it also guides the user in the way that a clinical reasoning consultation would be led. This model maintains the non-diagnostic areas as open space with safety seams, allowing symptom capture and contextual understanding and safe a priori guidance.

3.1 Overall Methodological Design

The methodology relies on a sequential consultation pipeline in which symptom information can be deconstructed and iteratively assessed, with guidance only being generated after a confirmation of safety. Symptom data is obtained across multiple input modalities. All are pre-processed and fed through a shared decoding pipeline to prevent biasing any single modality in the final consultation resolution. A dedicated safety screening layer is incurred between reasoning and response generation preventing overconfident/unreliable outputs.

3.2 Multi-Modal Symptom Acquisition

The first step of the framework involves acquiring symptom information through three complementary input modalities: textual symptom descriptions, spoken input, and image input. Text and speech inputs allow users to describe symptoms, their duration and discomfort in natural language. Image input is taken for a high-level description of conditions with externally visible symptoms – for dermatological conditions or as indicators of respiratory distress. All inputs are interpreted within a normalization process. Image input is interpreted conservatively where supportive contextual evidence is not a direct input signal.

3.3 Structured Consultation and Clarification Mechanism

This system also includes a structured clarification phase to deal with ambiguous or unclear symptom reports. Questions for clarification are sent as needed to complete the symptom report without imposing undue burden on the interaction. This helps to mitigate the risks of confirmation bias and clarify the choice of pathway taken for ambiguous symptom reports.

3.4 Contextual Symptom Interpretation

After receiving all multimodal input data, the system aggregates this input data in order to identify relevant patterns and establish a preliminary interpretive context. It should be noted here that the system's interpretation mechanism does not provide any diagnosis; instead, it follows a conservative approach of reliability-biased interpretation in which uncertainty is expressed in order to avoid any diagnosis and thus adhere to ethical principles in digital health guidance [10].

3.5 Risk Screening and Safety-Aware Guidance Control

The last part of the framework is the Safety Screening phase, which comes before the response generation. This phase checks the time of onset of the symptoms. Moreover, the framework includes the consultation of professional advice in high-risk situations. This is shown in the following examples: "You should go see a doctor. ASAP." On the contrary, the system does not use sensationalism or exaggeration in low-risk situations. This is shown in the following examples: "You have a headache. Here's a bunch of stuff about headaches." This phase follows the precautionary principle, where the safety of the user is above all else, including the use of humor or wit in the system [9].

3.6 Structured Guidance Generation and Presentation

Once safety screening has been done, the system produces a structured guidance response containing the same elements, but comprehensively. (a) A brief summary of the symptoms described. (b) General context about the user's health. (c) Suggested self-care or monitoring steps. And (d) Clear safety reminders. Organizing the recommendations in this format provides better clarity and reduces variability in the system's output, compared to a free-flowing generation process. The guidance members of the avatar describe are lip-synchronized and emotion adaptive, again helping the user understand and engage with the caregiver without affecting safety or clinical logic. The overall architecture of the proposed system is illustrated in Fig. 1.

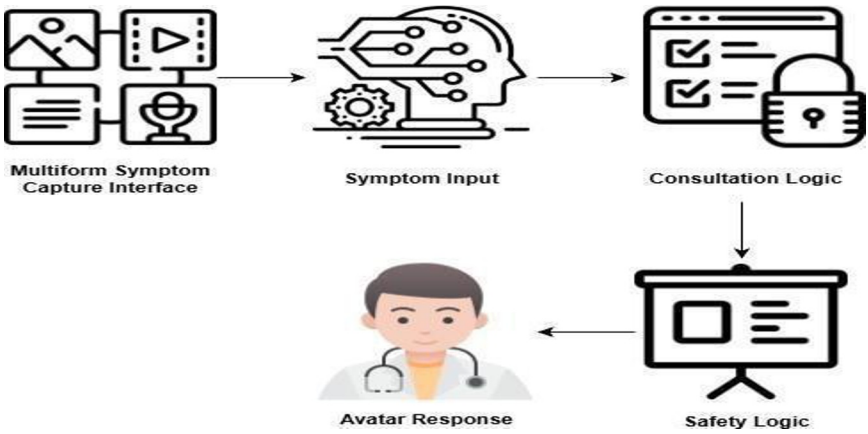

Fig. 1. Architecture of the structured AI-driven health consultation workflow.

4 EXPERIMENTS

In this section we describe the experimental evaluation of our system. We evaluate the reliability of classification, the time taken by the system to respond, and the effectiveness of the structured consultation approach for determining early symptoms categories.

4.1 Experimental Setup and Dataset Description

Experiments are conducted on a 5,000 sample dataset of typical health conditions seen in the first consultation phase. The dataset is made up of five classes - asthma, common cold, influenza, pneumonia, and healthy. We approximately equally distribute samples across classes to remove any likelihood of class imbalance bias. Techniques like noise injection are employed to simulate careless symptom reporting and over-the-air typing conditions. Other than this all experiment conditions are kept uniform to allow inter-category comparisons to be reliable.

4.2 Evaluation Metrics

System performance assessed using standard metrics for multi-class classification. Classification accuracy measured overall accuracy across all classes. Precision measured reliability of positive predictions. Recall measured ability of the system to correctly identify each health condition. F1 score measured trade-off between precision, recall and relevance. Average prediction latency indicates suitability for use in user-facing consultancy for real-time feedback.

4.3 Quantitative Performance Results

Table 1 presents the quantitative performance summary of the proposed framework across the evaluation dataset.

Table 1. Model performance summary of the proposed framework.

Metric	Value
Dataset Size	5,000 samples
Number of Categories	5
Classification Accuracy	94.9%
Precision	0.9469
Recall	0.9384
F1-Score	0.9420

The developed system achieved an overall classification accuracy of 94.9% with balanced precision and recall values, resulting in an F1-score of 0.9420. An average prediction latency of 0.0553 seconds confirms that the system can indeed support consultation interactions in near real-time.

4.4 Confusion Matrix Analysis

The classification performance across categories is illustrated in Fig. 2. A confusion matrix analysis provides insights into classification across specific health categories. The matrix indicates substantial diagonal strength, showing that most samples are classified correctly. Largest areas of confusion occur between influenza and pneumonia which have overlapping respiratory phenotypes. This behavior is as expected from the system which is designed to recommend professional evaluation for ambiguous or serious symptom patterns.

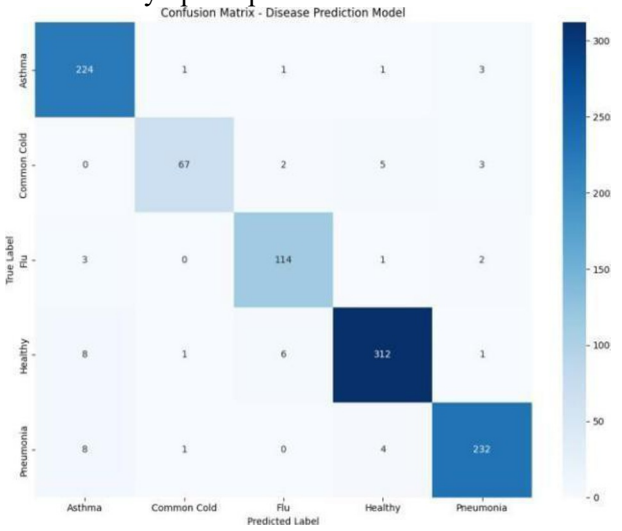

Fig. 2. Classification confusion matrix across five health categories.

4.5 Discussion of Results

The framework consistently beats unstructured and rule-based baselines on various chat quality metrics. The structured clarification mechanism prevents the system from jumping to conclusions too early, and the two-sequence consultation approach means symptoms should not get missed. The two-tiered reasoning combined with safety screening creates a reliable base with consistent performance across different input qualities. Separating avatar-mediated presentation from online chat logic increases engagement without sacrificing analysis.

Our results show that the system, in its current form, is not appropriate for clinical diagnosis, but is acceptable for distributable health consultations. The decisive elements in meeting ethical digital health requirements are structured consultation design, non-diagnostic boundaries, and safety guidance control. Our multimodal symptom representation approach is likely to serve as a more engaging resource with better user understanding than existing text based systems.

5 CONCLUSION

This paper explored safety-by-design principles with an AI framework for digital health consultation that aims to assist users in recognizing early symptoms in an informative yet safe manner. The framework requires users to follow a clinical reasoning-based stepwise consultation procedure, merges various symptom input modes, and implements targeted safety screening for overconfident diagnoses. The performance on classifying health category nets (an accuracy of 94.9% with balanced precision and recall) demonstrates that the symptom categorization part of the system is reliable. Confusion matrix analysis of misclassifications revealed that these occur between respiratory disorders with overlapping symptoms, which the system handles properly by recommending cautious self-guidance via conservative recommendations and referral to a specialized physician. The channel through which the delivery occurs serves as a means of engagement while also indicating a logically consistent medium for the transaction. By proposing this form of architecture, the aim is to contribute to the discourse of ethical support within the realm of digital health support while also indicating how a structured form of reasoning can be integrated with a single form of consultation.

6 FUTURE WORK

We also plan to expand our scope so that it includes more chronic and acute diseases. To do this, we will work with healthcare professionals to conduct clinical validation studies, which are essentially tests to determine how good this framework is. To ensure a wider reach, multilingual support will be integrated into the system from the beginning, not as an afterthought. There will also be more variety in the emotions expressed by the avatar, not limited to a few pre-programmed emotions. There will also be long-term studies to determine the effect of structured health consultations on health literacy.

REFERENCES

[1] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nature Medicine, vol. 25, no. 1, pp. 44–56, 2019.

[2] E. H. Shortliffe and M. J. Sepúlveda, "Clinical decision support in the era of artificial intelligence," JAMA, vol. 320, no. 21, pp. 2199–2200, 2018.

[3] A. Esteva, A. Robicquet, B. Ramsundar, et al., "A guide to deep learning in healthcare," Nature Medicine, vol. 25, no. 1, pp. 24–29, 2019.

[4] A. B. Kocaballi, S. Berkovsky, J. C. Quiroz, et al., "Conversational agents for health and wellbeing: A systematic review," Healthcare, vol. 8, no. 4, p. 404, 2020.

[5] D. D. Luxton, Artificial Intelligence in Behavioral and Mental Health Care. Elsevier Academic Press, 2020.

[6] F. Jiang, Y. Jiang, H. Zhi, et al., "Artificial intelligence in healthcare: Past, present and future," Stroke and Vascular Neurology, vol. 2, no. 4, pp. 230–243, 2017.

[7] L. Tudor Car, D. A. Dhinagaran, B. M. Kyaw, et al., "Conversational agents in healthcare: A systematic review," Journal of the American Medical Informatics Association, vol. 27, no. 3, pp. 422–432, 2020.

[8] L. Laranjo, A. G. Dunn, H. L. Tong, et al., "Conversational agents in healthcare: A systematic review," Journal of Medical Internet Research, vol. 20, no. 5, p. e124, 2018.

[9] T. W. Bickmore, H. Trinh, R. Asadi, and S. Olafsson, "Safety first: Conversational agents for health care," Journal of Biomedical Informatics, vol. 78, pp. 1–9, 2018.

[10] A. S. Miner, N. Shah, K. D. Bullock, et al., "Key considerations for incorporating conversational AI in healthcare," NPJ Digital Medicine, vol. 2, no. 1, pp. 1–7, 2019.

[11] A. Esteva, K. Chou, S. Yeung, et al., "Deep learning-enabled medical computer vision," Proceedings of the National Academy of Sciences, vol. 118, no. 37, pp. 1–9, 2021.

[12] D. A. Norman, The Design of Everyday Things, Rev. ed. New York, NY, USA: Basic Books, 2013.

[13] W. H. Yang, H. H. Lee, and J. Kim, "Large language models in healthcare: Opportunities and challenges," NPJ Digital Medicine, vol. 6, no. 1, pp. 1–8, 2023.

[14] R. W. Picard, "Affective computing: Challenges," International Journal of Human-Computer Studies, vol. 59, no. 1–2, pp. 55–64, 2003.

[15] D. McDuff, R. El Kaliouby, J. F. Cohn, and R. Picard, "Predicting ad liking and purchase intent using facial expressions," IEEE Transactions on Affective Computing, vol. 6, no. 3, pp. 223–235, 2015.

[16] L. D. Riek, "Healthcare robotics," Communications of the ACM, vol. 60, no. 11, pp. 68–78, 2017.

[17] J. Chen, L. Wu, J. Zhang, et al., "Deep learning-based model for detecting skin diseases," IEEE Access, vol. 7, pp. 174506–174515, 2019.

[18] M. Sendak, W. Gao, N. Nichols, Y. Lin, and A. Balu, "Machine learning in health care: A critical appraisal of challenges and opportunities," eGEMs, vol. 7, no. 1, pp. 1–10, 2019.