SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

HEALX: Agentic Health Monitoring with IoT, RAG, and Cloud-Native Multi-Agent AI

Achal Shirbhate[1], Saishwari Katkam[2], Disha Uplanchiwar[3], Dr. Santosh Warpe[4]

[1,2,3,4] Department of Computer Engineering, MIT Academy of Engineering, Alandi (D), Pune, India.

Abstract
HEALX is a cloud-based health monitoring system that uses wearable sensors, artificial intelligence, and medical knowledge retrieval to provide users with preliminary health-risk information. The system is based on a wearable device on ESP32 to collect vital physiological parameters such as heart rate, blood oxygen saturation (SpO2), and body temperature. These readings are sent to a backend based on FastAPI where they are processed along with the symptoms provided by the user. HEALX uses four specialised AI agents that provide a more structured process for health assessment: Symptom Interpreter, Symptom Summariser, Clarification Agent and Risk Assessment Agent. These agents work together, not unlike a single AI prompt. Language-based processing is done using Llama-3.3-70b-versatile via the Groq API. AI agents generate responses with relevant medical information provided by a Retrieval-Augmented Generation (RAG) mechanism using ChromaDB and the text-embedding-3-small model. One of the key features of HEALX is that the AI reasoning is not allowed to act independently of the sensor readings. Critical readings can bypass the normal AI reasoning process by applying programmatic safety rules to physiological values. The whole system is built with AWS services, such as a Virtual Private Cloud (VPC), Application Load Balancer, EC2 Auto Scaling, Amazon SQS, Redis, and Supabase PostgreSQL. React/Vite based web interface. Supports English, Hindi and Marathi. HEALX achieved an average end-to-end response time of 8.2 seconds, vector retrieval latency of <50 milliseconds, reported cloud uptime of 99.7%, and 90% risk-stratification accuracy in 20 validated cases in the reported simulations. Human annotators evaluated the follow-up questions, assigning them an average relevance score of 4.3 out of 5. The system is designed to help with preventive screening and health-related decision support and is not meant to replace professional medical diagnosis.

Keywords: Agentic AI, Health Monitoring, Internet of Things, Retrieval-Augmented Generation, Multi-Agent Systems, AWS, FastAPI, Llama-3.3, ChromaDB, Cloud Computing

I. Introduction

The use of wearable devices, cloud computing, and artificial intelligence has created new possibilities for remote health monitoring. Wearable devices can continuously collect physiological information, while cloud platforms make it possible to store and process this information. At the same time, artificial intelligence can help users understand symptoms and obtain preliminary health guidance.

However, most of the existing health-monitoring applications are concerned with collecting and displaying sensor data. Conversely, traditional symptom-checking apps typically depend on pre-established questions and fixed rules. The two approaches are frequently treated as separate systems. This makes it more difficult to link actual physiological measurements to user-reported symptoms and prior health data.

The proposed HEALX fills this gap by combining these components into one system. This project is an ESP32-based wearable sensing device and the cloud-based AI system. The wearable device makes physiological measurements in real time and the AI pipeline processes the symptoms of the user and combines them with the health information available.

HEALX is primarily designed to help users when they have no idea if their symptoms are trivial or should be treated by a doctor immediately. The system doesn't just display sensor values, it follows a complete workflow. It collects vital signs, interprets the user symptom description, queries a medical knowledge base for relevant information, asks additional questions when needed and ultimately generates a four-level risk assessment: Low, Medium, High or Emergency.

The system is designed as a screening and decision support platform, not as a diagnostic system. It is not intended to give formal medical diagnoses or prescriptions. Features such as medical-image analysis, multi-lead ECG analysis, live doctor consultation, and prescription generation are out of scope for the current project.

The major contributions of the HEALX project are:

1. Development of a low-cost wearable health-monitoring prototype.

2. Integration of four specialized AI agents for different stages of health assessment.
3. Use of Retrieval-Augmented Generation to provide relevant medical context.
4. Application of deterministic safety rules to physiological sensor values.
5. Deployment of the system using scalable AWS cloud infrastructure.
6. Development of a multilingual web interface supporting English, Hindi, and Marathi.
7. Generation of historical assessment records and shareable health reports.

II. Related work

The development of HEALX is based on several areas of existing research, including IoT-based health monitoring, AI-assisted clinical decision support, Retrieval-Augmented Generation, and multi-agent systems.

A. IoT-Based Health Monitoring

IoT has been widely used in healthcare to collect physiological information from sensors and transmit it to cloud platforms. These systems make remote monitoring possible without requiring users to manually record their measurements.

Nonetheless, most previous IoT health monitoring devices are mostly concerned with collecting information or issuing alerts. These devices collect the data and compare them with pre-set limits. While this is an efficient way of identifying abnormal data, it might prove difficult in situations where there are various factors to consider.

B. AI-Based Health Decision Support

AI is increasingly becoming a more useful tool in healthcare. The conventional healthcare applications used to be based on task-specific models and classifiers which would solve the particular problem at hand. Thanks to the emergence of LLMs, AI tools are capable of comprehending and handling complicated healthcare data expressed in natural language.

LLMs can be beneficial when the patient describes their symptoms independently, since the model can interpret the description and provide an appropriate reply. At the same time, there are certain drawbacks to depending solely on language models. For example, AI models can give false information due to hallucination. Moreover, the language model is unaware of the patient's current health status and physiologic data.

In order to address these shortcomings, HEALX leverages the capabilities of the language model in conjunction with medical data obtained using the RAG technique and physiological data obtained in real time using wearable devices. Through this, the health-risk assessment conducted by the AI is grounded not only on the user's input but also on additional medical data and physiological information.

III. Proposed Healx System

A. System Architecture

HEALX employs a three-layered architecture which includes a biometric sensing layer, an agentic AI backend layer, and a cloud infrastructure layer. The first layer is in charge of gathering the physiological information. An ESP32 microcontroller coupled with a MAX30102 pulse-oximeter and a DS18B20 temperature sensor are used in this process. The second layer consists of the FastAPI backend and the multi-agent AI workflow. This layer receives the sensor readings and the symptoms entered by the user.

The third layer contains the cloud infrastructure application. It handles traffic management, computing resources, data storage, caching, security, monitoring, and scaling. This separation allows each part of the system to perform a specific function while making the overall architecture easier to maintain and scale.

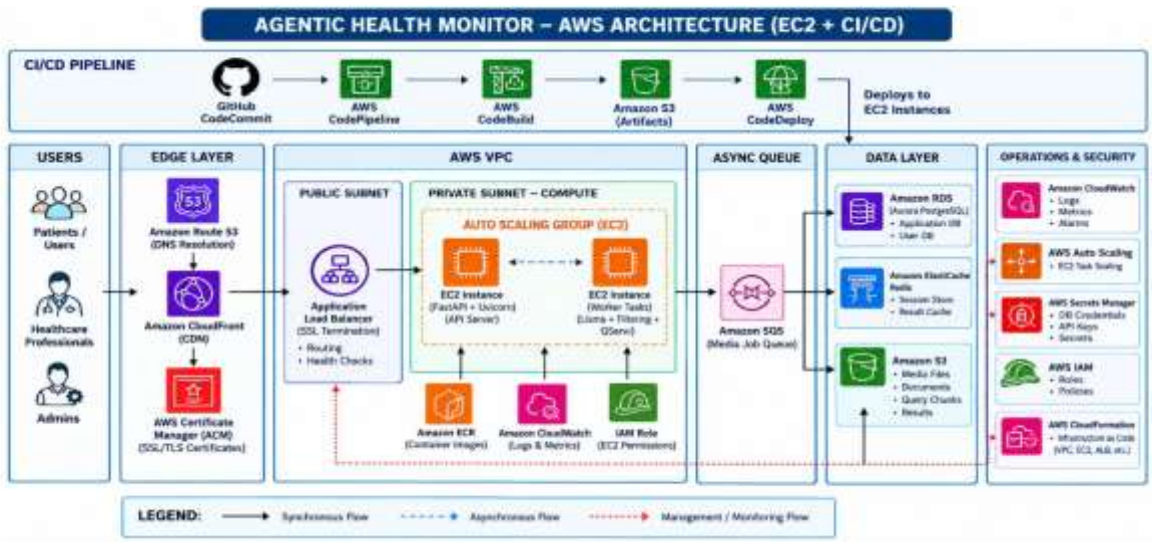

Fig. 1. AWS-based cloud architecture of the HEALX agentic health-monitoring system.

B. Asynchronous Cloud Processing

A major challenge in AI-based applications is that language-model processing can take longer than a normal web request. If the browser has to wait for every AI operation to finish, it can result in request timeouts or a poor user experience.

HEALX addresses this problem by using asynchronous processing.

When a request reaches the public API, the backend validates the data and places the longer-running AI task into an Amazon SQS FIFO queue. Worker EC2 instances then retrieve the queued tasks and execute the multi-agent reasoning process. An Application Load Balancer distributes incoming API requests, while EC2 Auto Scaling increases or decreases computing capacity according to system demand. This architecture separates the public API request from the longer AI-processing task and reduces the possibility of browser or gateway timeouts.

C. Data Persistence and Security

The structured information that HEALX requires is stored in the Supabase PostgreSQL database and includes user profiles, assessments, symptoms, vitals, matched conditions, and generated recommendations. Redis is used to keep temporary session information, rate-limiting counters, and agent execution information. Private archival storage for the generated clinical report and templates is done via Amazon S3. Security measures are provided by AWS VPC, HTTPS/TLS traffic, AWS Secrets Manager, and AWS IAM Roles. System monitoring is accomplished using CloudWatch. The deployment process is automated via AWS CodePipeline, CodeBuild, and CodeDeploy.

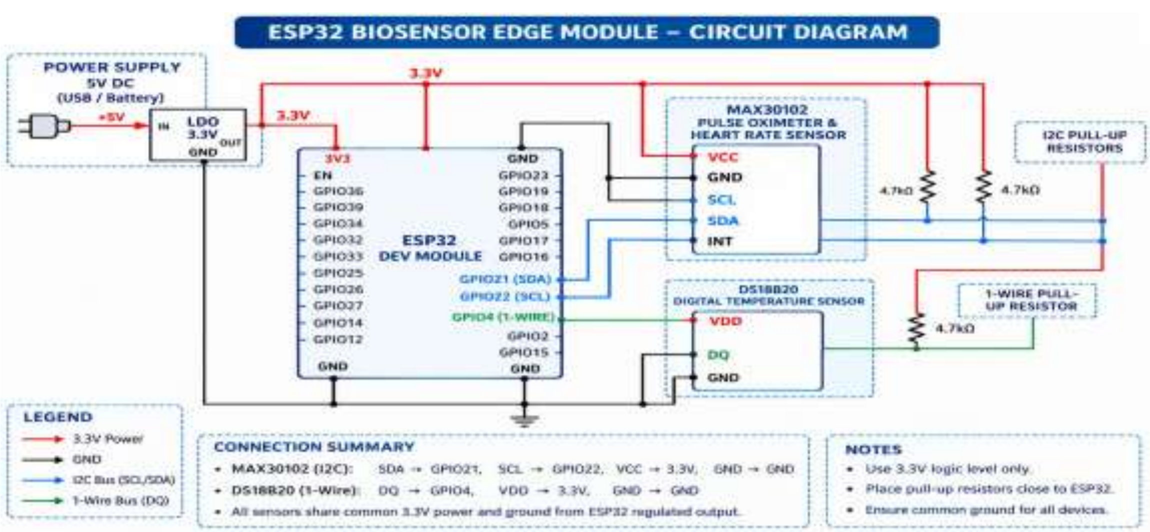

Fig. 2. Circuit diagram of the ESP32 biosensor edge module.

D. Multilingual Interface and Reporting

The HEALX platform has been designed to be easy to use for users of different languages. The dashboard displays patient information, symptoms, and physiological readings.

As soon as the assessment process is completed, a PDF file can be generated that is printable and shareable. Users can also generate an assessment URL link that is read-only, enabling access to the assessment data beyond the application itself.

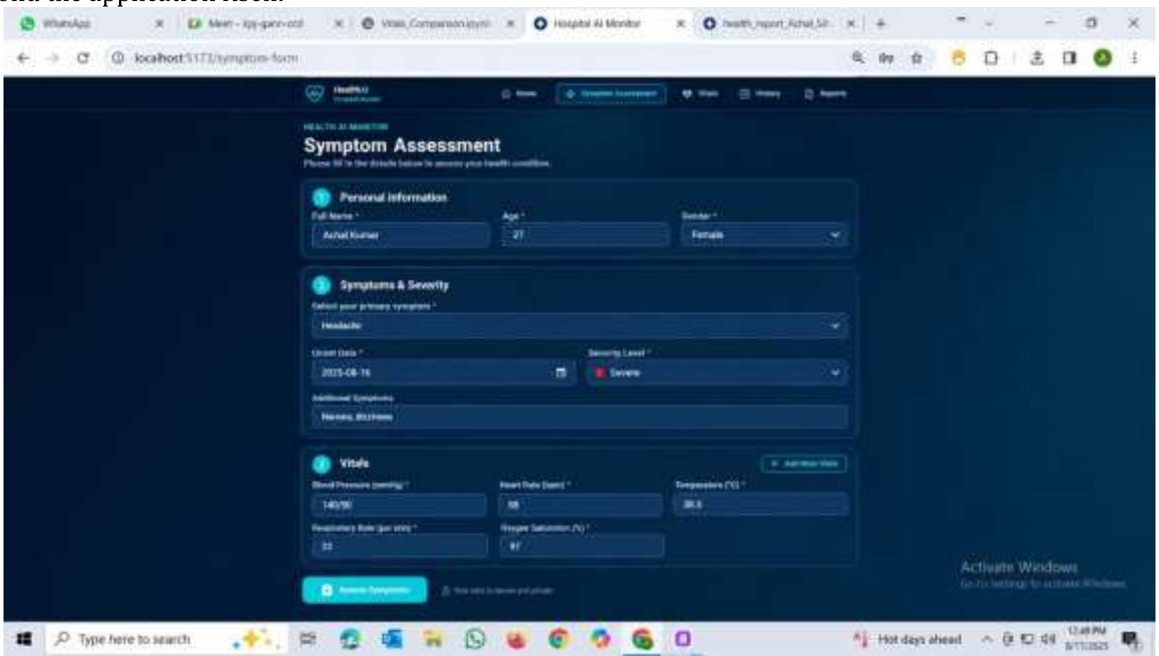

Fig. 3. HEALX Symptom Assessment web interface.

V. EXPERIMENTAL EVALUATION AND RESULTS

Evaluation of the HEALX software involved testing the program with simulated cases of patients with varying degrees of health risk. Criteria used for evaluation included the functionality of the system, the reaction time, the performance of the vector search, the reliability of the cloud system, the consistency of the multilingual interface, and the performance of the risk-stratification tool.

A. Low-Risk Case

The first example is a 22-year-old woman with symptoms such as a runny nose, sore throat, sneezing, and headaches.

The physical findings include:

- Temperature: 37.4 °C
- SpO2: 98%
- Heart rate: 78 BPM

The Symptom Interpreter found the patient's symptoms to be of low significance. As the information provided was deemed adequate, the clarification phase was bypassed. The RAG module provided two references for viral rhinitis. In the risk-assessment phase, the common cold was predicted to be the probable condition with a probability index of 0.82.

B. High-Risk Case

The second case is a 58-year-old male who experienced severe chest pain, described as crushing pain spreading from the chest toward the left arm, shortness of breath, profuse sweating, and high blood pressure.

The following readings were taken:

- Heart rate: 102 BPM
- SpO2: 94%

As the symptoms implied a possibility of a serious condition, the Symptom Interpreter automatically activated the emergency override function. The clarification step was intentionally skipped in order to speed up the process. Three articles concerning acute coronary syndrome were found using the RAG system. During the risk-assessment step, a reported risk score of 0.91 for acute myocardial infarction was obtained.

C. Performance Results

The performance figures obtained are summarized in Table I.

Performance Measure	Observed Result
Average full multi-turn execution latency	8.2 seconds
Single-agent response latency	1.1–1.8 seconds
Vector-search latency	Less than 50 ms
Risk-stratification accuracy	90% (18/20)
Follow-up question relevance	4.3 / 5.0
Reported cloud uptime	99.7%
English / Hindi / Marathi layout integrity	100%

The average latency of 8.2 seconds indicates that decoupling the public API and AI inference by asynchronous queuing prevented blocking of the browser during more time-consuming inference procedures. The vector-search time was consistently below 50 milliseconds, indicating that semantic searching did not significantly impact response latency. The system reached 90% accuracy in stratifying risk across 20 validated cases, with 18 cases classified correctly. Follow-up questions received a human relevance score of 4.3 out of 5. The evaluation also revealed a safety-oriented trade-off made by the system: because the classification threshold is set on the safe side, uncertainty may sometimes lead to unnecessarily high risk stratification. While this ensures that no dangerous cases are overlooked, the limited number of evaluated cases prevents any conclusions about clinical effectiveness.

VI. Discussion

Among the positive aspects of the system is its approach to splitting the health-assessment procedure into separate tasks. If a single AI prompt were responsible for symptom recognition, retrieval of medical knowledge, analysis of sensor readings, question formulation, risk assessment, and generation of the response, the prompt would be difficult to manage. In HEALX, these tasks are distributed among specialized agents, so each agent concentrates on a specific portion of the procedure and structured validation takes place between stages. This simplifies the flow and gives more control over the outputs produced at each stage.

Moreover, the RAG-based approach is preferable to fine-tuning a medical model, since it allows the medical reference material to be updated independently of the language model.

Another important aspect of the system is that the safety criteria are explicitly separated from the language model. This matter because safety criteria should not depend solely on how the language model interprets a sentence; explicit safety rules can therefore override less definite semantic understanding when necessary.

There are also a number of limitations of HEALX. First, the quality of the assessment depends on the accuracy and completeness of the description provided by the user, and the system may not evaluate incomplete input equally well. Second, the system relies on external services such as the Groq API, so API rate limits or other external factors could influence response time. Third, the wearable device could introduce measurement error due to movement, incorrect sensor placement, or communication issues.

It should also be noted that the evaluation was performed on a small set of simulated cases, so the reported accuracy and latency figures should be considered engineering-level evaluations rather than clinical validation.

VII. Conclusion And Future Scope

HEALX presents an end-to-end health-monitoring platform that connects wearable IoT sensors with a cloud-based multi-agent AI system. The system combines an ESP32 sensing device with FastAPI, asynchronous SQS-based processing, Llama-3.3 language-model agents, ChromaDB-based RAG, PostgreSQL storage, Redis session management, and AWS cloud infrastructure.

By combining these technologies, HEALX provides a single platform for collecting physiological information, understanding user-reported symptoms, retrieving relevant medical knowledge, asking additional questions, and producing a preliminary risk assessment. The reported implementation achieved an average end-to-end latency of 8.2 seconds, vector retrieval latency below 50 milliseconds, 90% risk-stratification accuracy across 20 validated cases, and 99.7% reported cloud uptime. These results demonstrate the technical feasibility of combining IoT sensing, RAG, multi-agent AI, and cloud infrastructure within a single health-screening workflow.

There are several potential directions to enhance HEALX in the future.

First, edge AI may be employed to reduce reliance on constant internet connectivity, enabling certain computations to be performed closer to the wearable sensor. Second, additional sensors may be incorporated — for example, ECG and blood-glucose measurement — to provide richer information about a user's physiological state. Third, HEALX may be extended to support direct telemedicine interactions between

clinicians and patients based on the generated assessment, and federated learning may be explored for improved AI modeling with less centralized collection of user data.
Long-term analytics and predictions represent another avenue for enhancement, where future versions would consider historical trends in addition to current readings. Finally, compliance with healthcare data standards and integration with health ecosystems such as FHIR, HealthKit, and Health Connect may be considered as potential enhancements. These approaches would help HEALX evolve from a screening tool into a comprehensive preventive health-monitoring system.

ACKNOWLEDGMENT
The authors would like to thank Dr. Santosh Warpe for his guidance and valuable support during the development of the HEALX project. The authors also acknowledge the Department of Computer Engineering, MIT Academy of Engineering, Alandi (D), Pune for their support in the design, implementation, testing, and deployment of the project.

References

[1] C. Zheng, C. Liu, Y. Xie, Y. Liu, and M. Ju, "Towards expert-level medical question answering with large language models," arXiv preprint arXiv:2305.09617, 2025.
[2] Amazon Web Services, "AWS Well-Architected Framework," 2025.
[3] FastAPI, "FastAPI Documentation," 2025.
[4] ChromaDB, "Chroma Documentation," 2024.
[5] Supabase, "Supabase Documentation," 2024.
[6] Groq, "Groq API Documentation," 2024.
[7] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, and V. Natarajan, "Large language models encode clinical knowledge," Nature, vol. 620, pp. 172–180, 2023.
[8] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems," arXiv preprint arXiv:2303.13375, 2023.
[9] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," arXiv preprint arXiv:2304.03442, 2023.
[10] World Health Organization, "Noncommunicable diseases," 2023.
[11] X. Yang, A. L. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, and J. Bian, "A large language model for electronic health records," npj Digital Medicine, vol. 5, no. 1, p. 194, 2022.
[12] A. Kumar and V. Ramesh, "Raspberry Pi-based portable health monitoring station with cloud integration," IEEE Sensors Journal, vol. 21, no. 8, pp. 9872–9881, 2021.
[13] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.
[14] S. Patel and R. Gupta, "Real-time health monitoring using IoT and cloud computing," International Journal of Biomedical Engineering and Technology, vol. 34, no. 2, pp. 112–128, 2020.
[15] P. Rajpurkar, A. Y. Hannun, A. Haghpanahi, C. Bourn, and A. Y. Ng, "Cardiologist-level arrhythmia detection with convolutional neural networks," arXiv preprint arXiv:1707.01836, 2020.