ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Impact Factor: 8.206

Volume 9, Issue 4, April 2026

www.ijmrset.com ijmrset@gmail.com +91 6381 907 438 +91 63819 07438

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Explainable AI-Based Real-Time Patient Deterioration Prediction System using Multimodal Health Data

Roselin Lourd, Ragul T, Subitsha S, Pramila D

Department of Artificial Intelligence and Data Science, Sri Manakula Vinayagar Engineering College, Puducherry, India

ABSTRACT: Remote healthcare has come a long way, but a fundamental gap still exists: most telehealth platforms are built around scheduled video calls, not around keeping a constant eye on the patient between those calls. When something goes wrong — a sudden drop in blood oxygen, an irregular heartbeat at 2 am — there is often no system in place to catch it early. This paper presents an Explainable AI-Based Real-Time Patient Deterioration Prediction System using multimodal health data designed to fill exactly that gap. Patients or their caregivers can upload vital signs — heart rate, blood oxygen saturation (SpO2), blood pressure, and body temperature — through a secure web API. Those readings flow into an AWS cloud pipeline where a machine learning model called Isolation Forest checks for unusual patterns in real time. If something looks wrong, a risk score is calculated and the system sends the right level of alert to the treating clinician — anywhere from a low-priority notification to a critical emergency flag. Before any scheduled consultation, a Generative AI module powered by Google Gemini reads through the last 24 hours of the patient's data and writes a plain-language briefing so the doctor walks in already knowing what to look for. The whole system runs on serverless AWS infrastructure, which means it scales automatically and keeps costs low. Early evaluation results suggest the system cuts down on the time it takes to catch a deteriorating patient, reduces alert overload for clinicians, and helps doctors make faster decisions when it matters most.

KEYWORDS: Telehealth, patient monitoring, anomaly detection, Isolation Forest, emergency triage, generative AI, AWS cloud, machine learning, healthcare IoT, clinical decision support

I. INTRODUCTION

Healthcare has changed a great deal in recent years. Wearable devices, faster internet connections, and a growing demand for remote services have pushed telehealth into the mainstream. Millions of patients now manage chronic conditions like diabetes, heart disease, and hypertension without ever needing to visit a clinic for routine check-ins. In theory, this is a win for everyone — patients save time, hospitals reduce congestion, and care reaches people in rural or underserved areas who might otherwise go without it.

In practice, though, most telehealth platforms still have a serious blind spot. They are good at connecting a patient with a doctor over a video call. What they are not good at is watching what happens to that patient in the hours and days between those calls. A person with a cardiac condition might have three dangerous arrhythmia episodes in the middle of the night, and unless they happen to mention it at their next appointment, no one will ever know. That delayed detection is not a minor inconvenience — it is a patient safety problem.

The system described in this paper was built to address that specific problem. Instead of waiting for a patient to report symptoms, it monitors uploaded vital signs continuously, uses machine learning to spot early warning signs, and notifies clinicians before a situation becomes a crisis. On top of that, it uses a large language model to write a short clinical briefing before each consultation — so the doctor does not have to spend the first five minutes of every appointment trying to piece together what has been happening with the patient.

The rest of this paper is organised as follows. Section II reviews existing research on remote monitoring, anomaly detection, and AI in clinical settings. Section III explains how the system is built and why specific design choices were

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|
International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)
(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

made. Section IV covers the technology stack. Section V discusses what the system is expected to achieve and how it will be evaluated. Section VI summarises the work and describes what comes next.

II. LITERATURE REVIEW

Before designing this system, we looked carefully at what has already been done in the three areas it touches: remote patient monitoring, machine learning-based anomaly detection, and AI-assisted clinical decision support. The table below summarises five key papers that shaped our thinking.

Table 1: Summary of Existing Research in Remote Monitoring, Anomaly Detection, and AI-Based Clinical Systems

No.	Paper Title	Author	Key Points	Limitation / Remark
1	Real-Time Health Monitoring Using IoT and Cloud	Firouzi et al., 2020	Proposes an IoT–cloud framework for continuous vital monitoring with rule-based alert generation.	Does not incorporate ML anomaly detection or AI-generated clinical summaries.
2	Machine Learning for Anomaly Detection in Patient Data	Goldstein & Uchida, 2016	Benchmarks unsupervised anomaly detection methods including Isolation Forest on health datasets.	Not integrated with live telehealth pipelines or clinician alert systems.
3	Clinical Decision Support Using AI	Topol, 2019	Shows how AI improves diagnostic accuracy through deep learning on imaging and structured records.	Focuses on diagnosis, not real-time remote triage or emergency alert generation.
4	Wearable IoT Sensors for Patient Monitoring	Dias & Paulo Silva Cunha, 2018	Reviews wearable sensors for SpO2, heart rate, and temperature; discusses data transmission protocols.	Clinical integration and data interpretation remain largely manual.
5	Generative AI in Healthcare Summarization	Thirunavukarasu et al., 2023	Reviews large language models for clinical note summarisation and patient-facing health communication.	Limited focus on pre-consultation briefing or emergency-specific triage.

The picture that emerges from this review is encouraging but incomplete. Researchers have built solid IoT monitoring platforms, proven that Isolation Forest works well on health data, and shown that AI can meaningfully support clinical decisions. What does not yet exist is a single system that ties all of these pieces together — one that collects vitals, detects anomalies, prioritises alerts, and then prepares the clinician for the conversation, all in real time. That integration is what this project attempts.

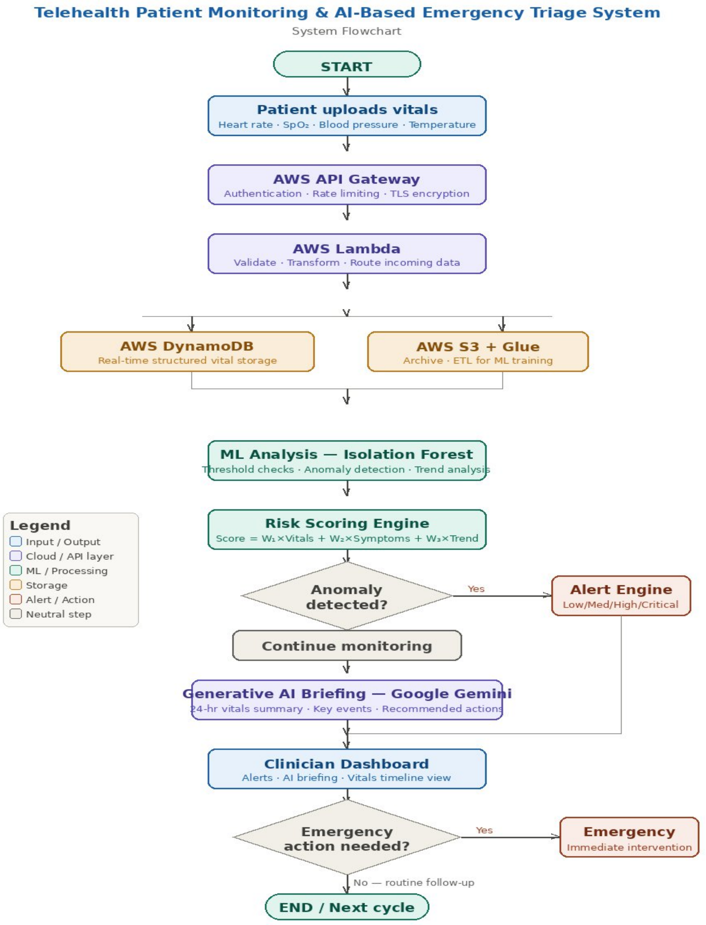
ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. PROPOSED SYSTEM ARCHITECTURE AND METHODOLOGY

The system is organised into five layers, each with a specific job. Together they form a continuous pipeline that starts the moment a patient submits a vital reading and ends when a clinician receives an actionable alert or a pre-consultation



briefing. Figure 1 below shows the complete flow.

Fig. 1: End-to-End System Flow of the Proposed Patient Deterioration Prediction System

A. Data Ingestion Layer: Everything starts when a patient or caregiver submits a set of vital sign readings. In the current version of the system this is done manually through a web form, but the architecture is designed so that IoT wearables can plug in directly once they are available. Before any data goes anywhere, AWS API Gateway checks the request — is the user authenticated? Does the request look legitimate? Is it coming in at a reasonable rate? Only if all of those checks pass does the data move forward. This is not just good security practice; it also prevents the downstream ML components from being flooded with junk data that could trigger false alerts.

B. Cloud Processing Layer: Once data clears the gateway, AWS Lambda takes over. Lambda is a serverless compute service, which in plain terms means that the system automatically spins up as many processing units as it needs to handle the current load — whether that is ten patients or ten thousand — and then scales back down when things are quiet. Lambda validates the incoming reading, formats it consistently, and sends it to two places at once: AWS DynamoDB for fast, structured storage that the system can query instantly, and AWS S3 for long-term archival. A separate AWS Glue job runs periodically to clean and prepare the stored data for retraining the ML model as new patient patterns accumulate.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

C. Machine Learning Analysis Layer: The ML module works in three stages, each catching things the previous one might miss. First, static threshold checks flag any reading that falls outside known safe ranges — a heart rate above 100 bpm at rest, an SpO2 below 94%, systolic blood pressure above 140 mmHg, and so on. These rules are simple but fast, and they catch the most obvious problems immediately. Second, an Isolation Forest model looks at the full set of readings together rather than each one in isolation. Isolation Forest works by trying to isolate each data point from all the others through random splits; points that are genuinely unusual get isolated quickly, while normal readings take many splits to separate. It is particularly good at catching subtle combinations — a heart rate that is only slightly elevated on its own but is paired with a temperature and blood pressure that together paint a worrying picture. To ensure explainability, the system highlights key contributing features (such as abnormal heart rate or SpO2 variations) responsible for anomaly detection, enabling clinicians to interpret the model's decisions effectively. Third, a sliding-window analysis looks at how each vital has been trending over the past few hours, catching slow deterioration that neither threshold checks nor the Isolation Forest would flag as unusual on a single reading.

The three stages feed into a risk scoring formula:

$$\textbf{Risk Score} = (\textbf{Vital Anomaly Score} \times W_1) + (\textbf{Symptom Severity} \times W_2) + (\textbf{Trend Deviation} \times W_3)$$

The weights W_1, W_2, and W_3 are calibrated on historical patient data. The final score sits between 0 and 100 and maps directly to one of four alert levels: Low (0–30), Medium (31–60), High (61–80), and Critical (81–100). A Critical score means the clinician should be notified immediately; a Low score means the reading is filed away quietly for the next scheduled review.

D. Alert Generation Engine: The alert engine takes the risk score and decides who needs to know, how urgently, and how often. One of the more important design decisions here was building in a cooldown mechanism: if a patient's score keeps bouncing between 55 and 65, the system does not send a separate Medium alert every five minutes. Alert fatigue — where clinicians become desensitised to notifications because there are simply too many of them — is a well-documented problem in hospital settings, and it is just as real in remote monitoring. The cooldown ensures that alerts stay meaningful.

E. AI Briefing Layer: The final layer is arguably the most distinctive feature of the system. Before a consultation, the briefing module pulls together everything that has happened with the patient in the last 24 hours: average values for each vital, any periods where the risk score was elevated, specific anomalous events detected by the ML module, and how recent readings compare to the patient's own historical baseline. This information goes into a carefully structured prompt that is sent to Google Gemini, which produces a short plain-language summary. The output looks something like this: three episodes of elevated heart rate between 1 and 3 am, SpO2 dropped below 93 percent twice, blood pressure trending upward over the past 48 hours — recommend reviewing current medication dosage. The briefing is formatted as structured JSON so it can be displayed cleanly in the clinician dashboard, but the language is deliberately human rather than technical, because the goal is to save the doctor time, not add another data-heavy screen to scroll through.

IV. TECHNOLOGY STACK AND IMPLEMENTATION

Python is the backbone of the system's analytical components. It was the natural choice given the quality of its data science ecosystem: Scikit-learn provides the Isolation Forest implementation, Pandas and NumPy handle time-series processing and statistical feature extraction, and the Google Generative AI library connects the system to Gemini.

On the cloud side, every component is managed by AWS. API Gateway handles inbound requests; Lambda runs the processing logic without needing a dedicated server; DynamoDB stores vitals with a composite key structure (patientId as the primary key, timestamp as the sort key) that makes time-ordered queries fast and efficient; S3 holds raw data and model training files; and Glue automates the ETL pipeline that keeps the ML model up to date. Access to every resource is controlled by IAM roles following the principle of least privilege — each component can only touch the specific resources it needs and nothing else. All data at rest is encrypted using AWS-managed keys.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. EXPECTED OUTCOMES AND PERFORMANCE EVALUATION

We expect the system to improve telehealth outcomes in five concrete ways. The first is simply continuous visibility: rather than only knowing how a patient was doing at their last appointment, clinicians will have access to a live picture of how that patient is doing right now and how things have been trending over recent days.
The second is earlier detection. The combination of threshold rules, Isolation Forest, and trend analysis should catch deteriorating conditions significantly earlier than a purely threshold-based system, because many health crises do not announce themselves with a single alarming reading — they unfold gradually across multiple vital signs over several hours.

The third is smarter alerts. By scoring and ranking events rather than treating every threshold breach as equally urgent, the system should meaningfully reduce the number of low-value notifications that reach clinicians. This matters because in real clinical practice, alert fatigue causes people to start ignoring notifications — which defeats the entire purpose of monitoring.

The fourth is time savings. Automated anomaly analysis removes a chunk of manual review work. In a typical telehealth workflow, a clinician might spend several minutes before each consultation manually scanning through recent vital uploads. The AI briefing condenses that into a 30-second read.

The fifth is decision quality. A doctor who walks into a consultation already knowing that the patient had three high-risk episodes in the past 48 hours is better positioned to ask the right questions and make better-informed choices than one who is finding out for the first time during the appointment.

To validate these claims, we plan to measure anomaly detection performance (precision, recall, and F1-score) against a labelled dataset of known clinical events, track the latency between vital submission and alert generation, and collect clinician feedback on the usefulness and accuracy of the AI briefings using standardised usability questionnaires. We will also compare detection accuracy against a baseline system that uses threshold rules only, to quantify the specific contribution of the Isolation Forest component.

VI. CONCLUSION AND FUTURE WORK

Remote patient monitoring has enormous potential, but that potential is currently limited by a lack of intelligent automation. Most systems today do a reasonable job of collecting data but a poor job of making sense of it quickly enough to matter. This project set out to close that gap by combining continuous vital sign monitoring, machine learning-based anomaly detection, priority-based alerting, and AI-generated clinical briefings into a single, coherent system.

What we built is a serverless, event-driven platform that watches for health deterioration as it happens, notifies the right people at the right level of urgency, and prepares clinicians to have better conversations with their patients. It is not a replacement for clinical judgment — it is a tool that gives clinicians the information they need to exercise that judgment more effectively and more quickly.

There is still a great deal of room to grow. The most important next step is moving from manual vital uploads to automated ingestion from IoT wearable devices, which would make monitoring truly continuous. Beyond that, we plan to explore more sophisticated predictive models — recurrent neural networks and transformer-based architectures that could forecast deterioration before it registers as an anomaly, not just after. A mobile application for patients to track their own readings would help with engagement and adherence. Full integration with hospital electronic health record systems would give the AI briefing module access to medication histories, previous diagnoses, and lab results, making its summaries significantly richer. And eventually, the system will need to meet HIPAA and applicable data protection standards before it can be used in a real clinical setting — work we consider essential and have already begun planning.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

[1] F. Firouzi, B. Farahani, M. Weinberger, A. DeFazio, and R. Forgeat, "Harnessing the Power of Smart and Connected Health to Tackle COVID-19," IEEE Internet of Things Journal, vol. 7, no. 11, pp. 10879–10897, 2020.
[2] M. Goldstein and S. Uchida, "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data," PLOS ONE, vol. 11, no. 4, 2016.
[3] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature Medicine, vol. 25, pp. 44–56, 2019.
[4] D. Dias and J. Paulo Silva Cunha, "Wearable Health Devices — Vital Sign Monitoring, Systems and Technologies," Sensors, vol. 18, no. 8, p. 2414, 2018.
[5] A. J. Thirunavukarasu et al., "Large language models in medicine," Nature Medicine, vol. 29, pp. 1930–1940, 2023.
[6] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation Forest," in Proc. 8th IEEE International Conference on Data Mining (ICDM), pp. 413–422, 2008.
[7] Amazon Web Services, "AWS Lambda Developer Guide," AWS Documentation, 2023. [Online]. Available: https://docs.aws.amazon.com/lambda/
[8] Google, "Gemini API Documentation," Google AI for Developers, 2024. [Online]. Available: https://ai.google.dev/
[9] World Health Organization, "Global Strategy on Digital Health 2020–2025," WHO, Geneva, 2021.
[10] B. Chowdhury and R. Khosla, "RFID-based Hospital Real-time Patient Management System," in Proc. 6th IEEE/ACIS International Conference on Computer and Information Science (ICIS), 2007.