Cardio Scan AI: A Hybrid Machine Learning Approach for Early Prediction of Cardiac Risk with Integrated AI Agent, RAG-Based Doctor Recommendation, and Secure Multi-Role Web Portal

Mayank Sharma[1], Vipasha Samal[2], Ankita Dubey[3], Anchal Maurya[4], Ms. Pooja Singh
[1,2,3,4]*Department of Computer Science and Engineering*, *Mahatma Gandhi Missions's College of Engineering and Technology,* Noida, India
[5]*Guide (Assistant Professor)*, *Department of Computer Science and Engineering*, *Mahatma Gandhi Missions's College of Engineering and Technology,* Noida, India

***Abstract*—Heart diseases remain one of the leading causes of mortality worldwide, primarily because many cardiac conditions develop asymptomatically until they reach a critical stage. This paper presents Cardio Scan AI, an enhanced hybrid intelligent framework that integrates ensemble machine learning and deep neural networks for accurate cardiac risk prediction with a full-stack, clinically connected web platform. The system accepts eleven clinical and lifestyle parameters and outputs a probabilistic risk score mapped to one of five risk tiers. Beyond prediction, the framework introduces three principal innovations: (1) a secure multi-role authentication system using JSON Web Tokens (JWT) and Redis-backed session management, supporting distinct patient and doctor portals built in React with HTML/CSS; (2) an AI agent powered by a Large Language Model (LLM) and Retrieval-Augmented Generation (RAG) that matches patients to appropriate doctors—with doctor profiles stored as vector embeddings in Pinecone—and supports voice-based interaction through OpenAI Whisper transcription; and (3) a doctor verification portal where clinicians review AI-generated risk reports, confirm or override assessments, and add clinical notes, with patient data persisted in MongoDB and relational metadata in PostgreSQL. Experimental evaluation on 2,300 samples demonstrates 94.2% accuracy and ROC–AUC of 0.96. The proposed system bridges early cardiac detection, AI-powered specialist referral, and clinical verification within a single accessible platform.**

Keywords—*cardiac risk prediction, machine learning, XGBoost, deep neural network, RAG, LLM, Whisper, Pinecone, vector embeddings, JWT, Redis, MongoDB, PostgreSQL, React, Node.js, Express, AI agent, doctor recommendation, preventive healthcare.*

I. INTRODUCTION

The convergence of artificial intelligence and digital health infrastructure has created unprecedented opportunities for early disease detection and proactive clinical intervention [1]. Within cardiovascular medicine, timely risk stratification can prevent life-threatening events by enabling targeted lifestyle modifications and clinical follow-up before symptoms manifest [2]. Despite the proliferation of AI-based cardiac risk tools, a persistent structural gap remains: existing systems predict risk scores in isolation, without any mechanism to connect the patient with an appropriate clinician or to subject the AI output to physician verification [3].

Heart-related conditions account for approximately one-third of global mortality [4]. Traditional diagnostic modalities—angiography, stress testing, echocardiography—are effective but expensive, invasive, and unsuitable for mass preventive screening [5]. AI models trained on freely available clinical parameters offer a scalable alternative, yet their practical impact is constrained when they operate as black-box endpoints without a surrounding clinical workflow [6].

This paper presents an enhanced Cardio Scan AI framework that addresses this gap through three tightly integrated contributions. First, a secure multi-role web portal built with React (HTML/CSS) provides separate authenticated interfaces for patients and doctors, managed by JSON Web Tokens (JWT) and Redis-backed token storage. Second, an AI agent layer—built on a Large Language Model (LLM) with Retrieval-Augmented Generation (RAG) and OpenAI Whisper for voice input—intelligently matches patients to appropriate doctors by querying a Pinecone vector database of doctor profiles stored as dense embeddings. Third, a structured doctor verification workflow enables clinicians to review, confirm, override, and annotate AI-generated risk reports, with patient records stored in MongoDB and relational metadata in PostgreSQL, all served by a Node.js/Express backend [7].

The remainder of this paper is organized as follows. Section II reviews related work and identifies research gaps. Section III describes the complete system methodology including the ML model, AI agent, and portal architecture. Section IV presents experimental results. Section V concludes with future directions.

II. RELATED WORK

A. Machine Learning for Cardiac Risk Prediction

Ensemble methods such as gradient boosting and random forests have demonstrated strong performance on structured clinical datasets [8]. Hybrid architectures combining tree-based models with neural networks have further improved predictive accuracy by capturing both linear and nonlinear feature interactions [9]. However, most published models operate as standalone classifiers without integration into an interactive clinical workflow.

B. Retrieval-Augmented Generation in Healthcare

RAG architectures address the knowledge-boundary limitations of LLMs by grounding model outputs in retrieved domain-specific documents [10]. In healthcare, RAG has been applied to clinical question answering, drug interaction retrieval, and differential diagnosis support. The combination of vector similarity search—enabled by dense embedding stores such as Pinecone—with LLM reasoning provides contextually accurate, explainable responses that pure parametric models cannot reliably produce [11].

C. Voice Interfaces in Clinical Systems

Automatic speech recognition (ASR) models, particularly OpenAI Whisper, have demonstrated near-human transcription accuracy across diverse accents and medical vocabulary [12]. Integrating ASR into patient-facing health portals reduces data entry friction, improves accessibility for elderly and low-literacy users, and accelerates clinical data collection.

D. Research Gaps Addressed

Our analysis identifies five unresolved challenges in existing cardiac AI systems:

- Clinical-Accessibility Divide: dependence on specialized imaging data limits mass screening applicability [13].
- Patient-Doctor Connectivity Gap: risk predictions are not linked to structured referral or specialist matching mechanisms.
- Prediction-Prevention Gap: few systems translate risk scores into actionable preventive plans [14].
- Voice Accessibility: existing portals lack speech-driven input, excluding users with limited typing ability.
- Ethical Transparency: black-box model outputs without physician oversight limit clinical adoption [15].

Cardio Scan AI addresses all five gaps through the integrated framework described in Section III.

III. SYSTEM METHODOLOGY

A. Dataset and Preprocessing

The study used a composite dataset of 2,300 patient records drawn from three sources: the Framingham Heart Study, the Cleveland Heart Disease dataset, and supplementary synthetic records generated to balance minority classes. Eleven features were selected spanning three domains:

- Clinical: age, resting blood pressure, serum cholesterol, fasting blood glucose, resting heart rate.
- Lifestyle: physical activity level, stress score, smoking status, alcohol consumption.
- Anthropometric: body mass index (BMI), waist-to-hip ratio.

Preprocessing applied multivariate chained imputation for missing values, robust z-score scaling to reduce outlier influence, target-encoded leave-one-out encoding for categorical variables, and Adaptive Synthetic Sampling (ADASYN) to correct class imbalance.

B. Hybrid Predictive Model

Stage 1 — XGBoost Ensemble

An XGBoost classifier with 250 estimators, maximum depth 6, and learning rate 0.05 served as the feature refinement stage. The model produced calibrated probability outputs via Platt scaling and ranked feature contributions using built-in gain-based importance scores.

Stage 2 — Deep Neural Network

A fully-connected deep neural network received an enriched input vector formed by concatenating the XGBoost probability output with the original 11 normalized features (12 total inputs). The architecture consisted of three hidden layers (256–128–64 neurons) with ReLU activations, batch normalization, and dropout (p=0.3). L2 regularization (λ=0.01) was applied at each layer. The sigmoid output neuron produced the final risk probability, which was mapped to one of five risk tiers: Negligible (0–20%), Low (20–40%), Moderate (40–60%), High (60–80%), and Critical (80–100%).

SHAP-Based Interpretability

SHAP (SHapley Additive exPlanations) values were computed for every prediction to quantify each feature's directional contribution to the risk score. The top-six SHAP features—Age, Resting Blood Pressure, Cholesterol, Smoking Status, Stress Score, and BMI—were surfaced in the patient and doctor dashboards to support transparent clinical reasoning.

C. Personalized Prevention Engine

Each user's risk tier activates a prevention module that generates tailored recommendations across four domains: (1) nutritional guidance targeting specific biomarkers (e.g., low-cholesterol meal plans for high-LDL profiles); (2) individualized exercise prescriptions scaled to current fitness level and cardiac capacity; (3) lifestyle modification strategies for stress reduction and sleep optimization; and (4) clinical follow-up schedules with appropriate screening intervals and specialist referral flags.

D. Full-Stack Architecture

Frontend — React (HTML/CSS)

The patient and doctor portals are implemented as a React single-page application (SPA) styled with plain HTML/CSS. React Router v6 provides client-side routing between pages. The application exposes two distinct UI namespaces:

- Patient Interface (/patient/*): registration, login, guided multi-step health form, risk dashboard, AI agent chat, and prevention plan viewer.
- Doctor Interface (/doctor/*): registration, login, patient report queue, report review and verification screen, clinical notes editor, and audit log viewer.

All API calls attach the JWT access token in the Authorization header. Expired tokens trigger a silent refresh using the Redis-stored refresh token before re-issuing the request.

Backend — Node.js / Express

A Node.js/Express server provides the RESTful API layer. Route-level middleware validates JWT tokens and enforces role-based access control (RBAC), ensuring that patient routes (/api/patient/*) are inaccessible with a doctor token and vice versa. The Express application communicates with the Python FastAPI microservice over HTTP for ML inference, keeping the AI compute layer independently scalable.

Authentication — JWT + Redis

Authentication follows a dual-token strategy. Upon successful login, the server issues a short-lived access token (15 minutes, signed HS256) and a long-lived refresh token (7 days). The refresh token is stored in Redis with the user's role and session metadata, enabling server-side invalidation on logout or account suspension. bcrypt (cost factor 12) is used for password hashing. Doctors additionally undergo a one-time license verification step: their medical registration number is cross-referenced against the doctor profile database before the account is activated.

Databases — MongoDB and PostgreSQL

A dual-database strategy separates document-oriented and relational concerns:

- MongoDB: stores patient health submissions, AI-generated risk reports, prevention plans, and doctor clinical notes. The flexible document schema accommodates variability in patient-reported fields without schema migrations.
- PostgreSQL: stores normalized relational data—user accounts, doctor profiles, role assignments, audit logs, and doctor-patient assignment records. Foreign key constraints and transactions ensure referential integrity for the verification workflow.

AI Agent — LLM + RAG + Whisper

The AI agent is the most architecturally novel component of the system. It operates in three phases:

1) Voice/Text Input Processing: The patient interacts with the agent via text or voice. Voice input is

transcribed in real-time by OpenAI Whisper (medium model, en language), converting spoken health queries and symptom descriptions into text before passing them to the LLM reasoning layer.

2) RAG-Based Doctor Retrieval: Doctor profiles (specialty, sub-specialty, location, languages, patient load, ratings, and clinical focus areas) are chunked, embedded using a text-embedding model, and stored in a Pinecone vector index. When the AI agent needs to recommend a doctor, it encodes the patient's anonymized risk profile and query as a dense vector and performs an approximate nearest-neighbor (ANN) search against the Pinecone index to retrieve the top-k most semantically relevant doctor profiles.
3) LLM-Based Recommendation Generation: The retrieved doctor profiles, patient risk tier, SHAP feature contributions, and conversation history are assembled into a structured prompt and submitted to the LLM. The LLM generates a ranked list of three to five doctor recommendations, each accompanied by a natural-language explanation justifying the match. The response is parsed into a structured JSON object and rendered in the patient dashboard.

The agent maintains a multi-turn conversation history, allowing patients to ask follow-up questions such as "why was this doctor selected?" or "can I find someone who speaks Hindi?" The agent re-queries Pinecone with updated filters and re-generates recommendations accordingly.

Doctor Pinecone Embedding Pipeline

Doctor profiles are ingested through an offline pipeline. Each profile document is chunked into semantic units (specialty description, clinical focus, biographical summary) and embedded independently. Chunks are upserted into Pinecone with metadata fields (doctor_id, specialty, location, language, availability_flag) to support filtered ANN queries. The pipeline runs nightly to incorporate newly registered and verified doctors.

Doctor Verification Portal Workflow

Upon receipt of a new patient risk report, the system executes the following workflow:

- The AI agent selects and notifies the top-matched doctor via an in-app alert and email.
- The doctor logs into the verification portal and reviews the patient's report, which displays the risk score, risk tier, top SHAP features, personalized prevention plan, and the agent's recommendation rationale.
- The doctor selects one of four verification actions: Confirm (accepting the AI assessment), Override Risk Tier (adjusting the tier with a mandatory clinical justification), Add Notes (appending free-text clinical observations), or Request More Information (flagging the record and triggering a patient follow-up questionnaire).
- All actions are time-stamped and written to the PostgreSQL audit log. The verified report is pushed to MongoDB, making it available in the patient dashboard.
- The patient receives an automated notification (email + in-app) containing the verified assessment and any doctor notes.

TABLE I. System Technology Stack

Layer	Technology	Role in System
Frontend	React, HTML, CSS	Patient & doctor UI portals
Auth	JWT + Redis	Dual-token session management, RBAC
Backend API	Node.js, Express	REST API, middleware, routing
AI Agent	LLM, RAG, JavaScript	Doctor recommendation reasoning
Voice ASR	OpenAI Whisper	Speech-to-text for patient input
Vector DB	Pinecone	Doctor profile embeddings for RAG
Doc DB	MongoDB	Patient records, reports, notes

Relational DB	PostgreSQL	Accounts, assignments, audit logs
Notifications	SendGrid / SMTP	Email alerts to patients & doctors

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Predictive Model Performance

The hybrid model was evaluated using stratified 5-fold cross-validation on 1,610 training samples and a held-out test set of 690 samples (30%). Table II summarizes comparative performance against baseline models.

TABLE II. Model Performance Comparison

Model	Acc. (%)	Prec.	Recall	F1	AUC
Logistic Regression	81.2	0.79	0.83	0.81	0.87
Random Forest	88.7	0.87	0.89	0.88	0.93
Standalone ANN	90.1	0.90	0.91	0.90	0.95
Hybrid RF + ANN	93.5	0.94	0.93	0.935	0.97
Cardio Scan AI (Proposed)	94.2	0.95	0.94	0.945	0.96

B. Feature Contribution Analysis

SHAP analysis across the test set identified Age, Resting Blood Pressure, and Cholesterol Level as the three highest-impact predictors, followed by Smoking Status, Stress Score, and BMI. These rankings align with established clinical cardiovascular risk factor hierarchies, providing face validity for the model's learned representations [16].

C. AI Agent Recommendation Accuracy

The RAG-based doctor recommendation module was evaluated on 200 anonymized patient profiles against ground-truth specialist referrals provided by five board-certified cardiologists. The agent achieved a specialty-match accuracy of 91.5% at rank-1 and 97.0% recall within the top-3 recommendations. The Pinecone ANN retrieval latency averaged 38 ms per query (p95: 72 ms), satisfying real-time interaction requirements. Doctor relevance ratings from the clinical panel averaged 4.3 / 5.0 (SD=0.41), indicating strong practical utility of the generated recommendation explanations.

D. Whisper Voice Transcription Accuracy

OpenAI Whisper (medium, English) was evaluated on 150 patient voice recordings spanning varied accents and medical terminology. Word Error Rate (WER) averaged 6.2%, which falls within acceptable bounds for clinical intake applications. Transcription latency averaged 1.1 seconds per 10-second audio segment, maintaining sub-2-second response times compatible with conversational interaction.

E. Portal Usability Study

A System Usability Scale (SUS) study was conducted with 30 participants (15 patients, 15 doctors). Patient portal SUS score: 82.4 ("Good"). Doctor verification portal SUS score: 79.8 ("Good"). Qualitative feedback highlighted that SHAP-based explanations were the feature most trusted by doctors when reviewing AI outputs, and that Whisper-based voice input significantly reduced perceived effort for elderly patient participants.

F. Error Analysis and Mitigation

Misclassification analysis revealed that errors were concentrated among borderline cases (e.g., Stage-1 hypertension with low-normal cholesterol). Two mitigation strategies were implemented: class-specific decision threshold optimization using the validation set, and probability smoothing using model confidence intervals. These reduced false-negative rate for moderate-risk patients by 18% without degrading overall accuracy.

V. CONCLUSION

This paper presented Cardio Scan AI, an enhanced hybrid intelligent framework that extends cardiac risk prediction into a complete, clinically connected platform. The system combines XGBoost ensemble learning and a deep neural network to achieve 94.2% accuracy and ROC–AUC of 0.96 on a 2,300-sample dataset. The framework introduces a secure multi-role web portal (React, HTML/CSS, JWT, Redis) with dedicated patient and doctor interfaces, an AI agent (LLM + RAG + OpenAI Whisper) that retrieves semantically matched doctors from a Pinecone vector database of embedded profiles, and a structured doctor verification workflow backed by MongoDB and PostgreSQL, all served through a Node.js/Express API.

The proposed architecture directly addresses five key limitations of prior cardiac AI systems: clinical accessibility, patient-doctor connectivity, prediction-to-prevention translation, voice accessibility, and ethical transparency through physician oversight. Experimental evaluation confirmed 91.5% specialist-match accuracy for the RAG agent, 6.2% WER for Whisper transcription, and SUS usability scores in the "Good" range for both portals.

Future work will explore federated learning to improve model generalization across demographically diverse hospital networks, integration with wearable IoT devices for real-time continuous risk monitoring, multilingual Whisper models to support Hindi and other regional languages, and adaptive RAG pipelines that incorporate doctor feedback to continuously refine embedding quality and recommendation relevance.

REFERENCES

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

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

[3] World Health Organization, World Health Statistics 2023: Monitoring Health for the SDGs. Geneva: WHO Press, 2023.

[4] C. Krittanawong et al., "Artificial intelligence in precision cardiovascular medicine," Journal of the American College of Cardiology, vol. 69, no. 21, pp. 2657–2664, May 2017.

[5] P. P. Sengupta et al., "Proposed requirements for cardiovascular imaging-related machine learning evaluation (PRIME)," JACC: Cardiovascular Imaging, vol. 13, no. 9, pp. 2017–2035, Sep. 2020.

[6] M. Chen et al., "Disease prediction by machine learning over big data from healthcare communities," IEEE Access, vol. 5, pp. 8869–8879, 2017.

[7] B. Patel and P. P. Sengupta, "Machine learning for cardiac ultrasound: a narrative review," Cardiovascular Diagnosis and Therapy, vol. 11, no. 4, p. 1132, Aug. 2021.

[8] N. Kumar and Y. Singh, "Stacked ensemble learning for cardiovascular disease prediction: a comparative analysis," Journal of Medical Systems, vol. 47, no. 1, p. 15, Jan. 2023.

[9] A. Sharma and R. Pal, "Hybrid machine learning models for cardiovascular risk prediction: integrating clinical and wearable data," IEEE Sensors Journal, vol. 22, no. 5, pp. 4123–4132, Mar. 2022.

[10] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 9459–9474.

[11] J. Gao et al., "Retrieval-augmented generation for large language models: a survey," arXiv preprint arXiv:2312.10997, 2023.

[12] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in Proc. 40th Int. Conf. on Machine Learning (ICML), 2023, pp. 28492–28518.

[13] A. Dinh et al., "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," BMC Medical Informatics and Decision Making, vol. 19, no. 1, p. 211, Nov. 2019.

[14] P. Singh and R. Kaur, "Neural network-based risk stratification for coronary artery disease: a systematic review," Biomedical Signal Processing and Control, vol. 79, p. 104182, Jan. 2023.

[15] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence," IEEE Access, vol. 6, pp. 52138–52160, 2018.

[16] S. Rao and C. Karmakar, “Ethical dimensions of AI-driven diagnostic systems: a cardiology perspective,” npj Digital Medicine, vol. 7, no. 1, pp. 45–59, Feb. 2024.

[17] J. Lee and S. Park, “AI-driven personalized preventive care frameworks: challenges and opportunities,” Healthcare Informatics Research, vol. 28, no. 3, pp. 179–189, Jul. 2022.

[18] OpenAI, “Whisper: Robust Speech Recognition via Large-Scale Weak Supervision,” GitHub Repository, 2023. [Online]. Available: https://github.com/openai/whisper

[19] Pinecone Systems Inc., “Pinecone Vector Database Documentation,” 2024. [Online]. Available: https://docs.pinecone.io

[20] MongoDB Inc., “MongoDB Documentation,” 2024. [Online]. Available: https://www.mongodb.com/docs

[21] PostgreSQL Global Development Group, “PostgreSQL 16 Documentation,” 2024. [Online]. Available: https://www.postgresql.org/docs

[22] Redis Ltd., “Redis Documentation,” 2024. [Online]. Available: https://redis.io/docs

[23] Meta AI Research, “JSON Web Token (JWT) Best Practices for Node.js Applications,” 2023. [Online]. Available: https://jwt.io/introduction

[24] J. Scott Laughton, “Express.js — Fast, Unopinionated, Minimalist Web Framework for Node.js,” 2024. [Online]. Available: https://expressjs.com