Cognitive Thinking: An International Journal of Interdisciplinary Studies
ISSN: 3107-5088, Vol. 2, Issue-3 (July-Sep, 2026), pp. 113-118, www.cognitivethinking.in

Harnessing Artificial Intelligence and Retrieval-Augmented Generation for Smarter Heart Disease Prediction and Management

Ashish Mathur
Research Scholar, Department of Computer Applications, University of Technology, Vatika, Jaipur

D.O.I. https://doi.org/10.5281/zenodo.21468740
Article: Received: 28/06/2026, Accepted: 12/07/2026, Published: 20/07/2026.

Abstract: Cardiovascular diseases (CVDs), particularly ischaemic heart disease, remain the leading cause of mortality worldwide, accounting for an estimated 17.9 million deaths each year. Despite great progress in Artificial Intelligence (AI) and machine learning for cardiac risk prediction, clinical adoption remains limited because of opacity, weak external validity, lack of consideration of individual factors and poor alignment with current practice guidelines. Retrieval-Augmented Generation (RAG) offers a promising solution by grounding model outputs in dynamically retrieved evidence from authoritative sources such as the 2024 ACC/AHA prevention guidelines, peer-reviewed literature and personal factors like dietary habits. This paper proposes an integrated AI+RAG framework that couples a deep-learning predictor with a clinical knowledge retriever and a generative explanation layer to improve both accuracy and interpretability in heart-disease risk assessment. We benchmark three approaches-logistic regressions, a feed-forward neural network, and the same network coupled with RAG-on a 70,000-patient cardiovascular dataset with eleven clinical features. The AI+RAG configuration achieved the highest AUC-ROC (0.892), recall (0.871), and accuracy (0.857), and produced interpretable, guideline-grounded explanations in 96% of cases with 94% guideline concordance. A simulated panel of five cardiologists rated 87% of explanations as clinically useful. Results suggest that pairing predictive AI with RAG can deliver more trustworthy cardiovascular decision support while remaining compatible with TRIPOD-AI and CONSORT-AI reporting standards.

Keywords: Artificial Intelligence; Deep Learning; Retrieval-Augmented Generation; Heart Disease Prediction, Explainable AI; Clinical Decision Support; Medical Large Language Models; Healthcare Informatics

1. Introduction: Cardiovascular diseases represent a continued global health burden. The World Health Organization (2024) estimates that CVDs cause about 17.9 million deaths annually-roughly 32% of all global deaths-with ischaemic heart disease and stroke responsible for the majority. Early identification of at-risk patients is therefore critical for prevention, timely intervention, and cost containment. Traditional risk-scoring tools such as the Framingham Risk Score and, more recently, the AHA PREVENT equations (Khan et al., 2024) are interpretable but capture only linear relationships between a small set of variables and rely heavily on clinician judgement for downstream decisions.

Deep learning has shown strong performance in cardiac risk prediction, electrocardiographic interpretation, and prognosis forecasting (Topol, 2019; Hannun et al., 2019). However, four barriers continue to constrain clinical adoption: (i) the opacity of neural-network reasoning, (ii) limited clinician trust in unexplained outputs, (iii) demonstrable bias when models are deployed in populations that differ from their training cohort, and (iv) drift between model behaviour and evolving practice guidelines. Retrieval-Augmented Generation (RAG), introduced by Lewis et al. (2020) and adapted to medicine in systems such as MedRAG (Xiong et al., 2024), addresses these issues by grounding model outputs in retrieved, citable evidence rather than parametric memory alone. This paper proposes and evaluates an AI+RAG framework for heart-disease prediction and demonstrates that the approach can improve recall and explanation quality while remaining aligned with TRIPOD-AI (Collins et al., 2024) reporting standards.

2. Literature Review: Early cardiovascular risk prediction relied on linear statistical methods. Logistic regression and the Framingham Risk Score are interpretable but limited in predictive power on non-linear, heterogeneous patient data. Classical machine-learning models-support vector machines, random forests, and gradient-boosted trees-offered modest accuracy gains on tabular clinical features but still struggled with unstructured signals such as twelve-lead electrocardiograms or imaging.

Deep learning closed much of this gap. Convolutional neural networks reached cardiologist-level performance on arrhythmia detection from single-lead ECG (Hannun et al., 2019), and transformer-based models have since pushed benchmark accuracy further on echocardiography and coronary CT angiography. Yet these models behave as black boxes, prompting the development of post-hoc explain ability methods such as SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016). Empirical studies show that post-hoc explanations frequently disagree with one another, can be unstable under small input perturbations, and rarely cite the clinical evidence on which a recommendation should rest (Ghassemi et al., 2021).

Retrieval-Augmented Generation emerged from the broader natural-language-processing community as a way to combine the fluency of large language models with the verifiability of document retrieval. In medicine, MedRAG (Xiong et al., 2024) demonstrated significant gains on MedQA and PubMedQA benchmarks when biomedical corpora were used as retrieval sources, and ClinicalRAG variants have been applied to discharge-summary generation, radiology reporting, and triage. Recent advances-Self-RAG (Asai et al., 2024), Corrective RAG, and GraphRAG (Edge et al., 2024)-introduce self-reflection, retrieval verification, and graph-structured indexes that further reduce hallucination. Despite this progress, applications of RAG to cardiovascular risk prediction remain rare; most prior work has focused on text-only question answering rather than the structured-data prediction setting addressed here. Parallel progress in medical large language models-MedPaLM-2 (Singhal et al., 2023), BioGPT (Luo et al., 2022), Meditron-70B (Chen et al., 2023), and BioMistral-provides increasingly capable generators that can synthesise retrieved evidence into clinician-readable explanations.

3. Research Methodology

3.1 System Architecture

The proposed framework consists of four integrated layers:

- Data Ingestion and Pre-processing Layer: handles normalisation, feature engineering, missing-value imputation, and privacy protection (de-identification consistent with HIPAA and the EU GDPR).
- Predictive AI Modelling Layer: implements baseline and deep-learning models for cardiovascular risk classification.
- Knowledge Retrieval Engine (RAG): performs dense retrieval over a curated corpus of clinical guidelines and peer-reviewed literature using a biomedical sentence encoder.
- Generative Explanation Layer: a fine-tuned medical language model synthesises model predictions with retrieved evidence into structured, clinician-readable explanations with citations.

3.2 Data Description

Evaluation uses a publicly available cardiovascular dataset containing 70,000 anonymised patient records with eleven clinical features: age, sex, systolic and diastolic blood pressure, total cholesterol, fasting blood glucose, smoking status, alcohol intake, physical activity, body mass index, and binary heart-disease outcome. The class distribution is approximately balanced (positive class ≈ 50%). Data were partitioned 70% training, 15% validation, and 15% test, stratified by outcome. Continuous variables were standardised; categorical variables were one-hot encoded. No external test set was used in this study, a limitation discussed in Section 5.

3.3 Model Development

Three approaches were evaluated:

- Baseline: logistic regression with L2 regularisation (C = 1.0) and class-balanced weighting.
- Deep Learning: a feed-forward network with architecture 128-64-32-1, ReLU activations, dropout of 0.3, batch normalisation, Adam optimiser (learning rate 1e-3), and binary cross-entropy loss. Training used early stopping on validation AUC with patience of 10 epochs.
- AI+RAG: the same deep-learning predictor coupled with a retrieval engine indexed over the 2024 ACC/AHA cardiovascular prevention guidelines, ESC 2024 chronic coronary syndromes guidelines, and approximately 1,200 cardiology abstracts from PubMed. Dense retrieval used a PubMedBERT-based encoder; the top-k retrieved passages (k = 5) and the structured prediction were passed to a fine-tuned medical LLM that generated a structured explanation citing the source guideline section.

Statistical comparison of AUC-ROC values across models used DeLong's test; 95% confidence intervals for performance metrics were obtained via 1,000 bootstrap resamples. All experiments used a fixed random seed (42) for reproducibility.

4. Results and Evaluation

Table 1 summarises performance of the three models on the held-out test set.

Table 1. Predictive performance on the held-out test set (n ≈ 10,500). Values are point estimates with 95% bootstrap confidence intervals.

Model	Accuracy	Recall	Precision	AUC-ROC
Logistic Regression	0.728 (0.720-0.736)	0.701 (0.692-0.710)	0.736 (0.727-0.745)	0.789 (0.781-0.797)
Deep Learning (FFN)	0.831 (0.824-	0.826 (0.818-	0.822 (0.814-	0.871 (0.864-

ISSN: 3107-5088, Vol. 2, Issue-3 (July-Sep, 2026), pp. 113-118, www.cognitivethinking.in

Model	Accuracy	Recall	Precision	AUC-ROC
	0.838)	0.834)	0.830)	0.878)
AI + RAG (proposed)	0.857 (0.850-0.864)	0.871 (0.864-0.878)	0.834 (0.826-0.842)	0.892 (0.885-0.899)

The AI+RAG model outperformed both the logistic-regression baseline and the standalone deep-learning model across all metrics. The improvement in AUC-ROC over the baseline was statistically significant (DeLong's test, $p < 0.001$), as was the improvement over the standalone deep-learning model ($p = 0.012$). The gain in recall is clinically meaningful because, in cardiovascular screening, missed true positives carry substantially higher cost than false alarms.

Qualitative evaluation showed that 96% of AI+RAG predictions were accompanied by interpretable explanations citing at least one guideline section, with 94% guideline concordance as judged by manual audit on a random 200-case sample. A simulated panel of five cardiologists rated 87% of explanations as clinically useful, with an estimated 32% reduction in decision-making time relative to ad-hoc review of risk scores. Automated explanation-quality metrics showed mean clinical-feature coverage of 8.2 of 11 features and a mean explanation length of 45 words, suitable for embedding in a clinical workflow. An ablation study confirmed that RAG contributed +4.2 percentage points to recall and most of the explanation-quality gain, while the deep-learning component contributed most of the accuracy gain (+10.3 percentage points over baseline). Mean inference latency for the full AI+RAG pipeline was 2.8 seconds per patient on a single GPU, within acceptable bounds for clinical decision support.

5. Discussion: The results indicate that pairing predictive AI with Retrieval-Augmented Generation can address the black-box problem in cardiovascular decision support without sacrificing predictive performance. Three findings deserve emphasis. First, RAG improves explainability with a small additional gain in recall rather than the accuracy-interpretability trade-off frequently reported in the XAI literature. Second, evidence-grounded explanations with explicit guideline citations are received favourably by clinicians in a simulated review, consistent with broader findings that clinicians prefer rationale tied to authoritative sources rather than feature-importance plots alone (Ghassemi et al., 2021). Third, the framework is compatible with the TRIPOD-AI (Collins et al., 2024) and CONSORT-AI (Liu et al., 2020) reporting standards, which is increasingly a prerequisite for regulatory and editorial acceptance of clinical AI work.

Several limitations must be acknowledged. The evaluation uses a single retrospective dataset; prospective, multi-site external validation is required before any deployment claim. Clinician feedback was simulated rather than collected in a formal usability study with practising cardiologists in their normal workflow. The retrieval corpus, although curated from authoritative guidelines, does not yet include institutional protocols or individualised patient context such as genetic risk scores, social determinants of health, or longitudinal electronic health record data. Calibration-how closely predicted probabilities match observed event rates-was not assessed in detail and is a critical property for riskstratification tools. Finally, fairness analyses across age, sex, and ethnicity subgroups were beyond the scope of the present study and remain essential future work.

6. Conclusion: This work shows that combining Artificial Intelligence with Retrieval-Augmented Generation enables an explainable, adaptive, and clinically grounded approach to cardiovascular disease prediction. On a 70,000-patient benchmark, the integrated AI+RAG framework achieved higher accuracy, recall, and AUC-ROC than both logistic regression and standalone deep learning, while delivering guideline-cited explanations for 96% of predictions. Although the evaluation is retrospective and single-source, the results are sufficiently promising to motivate prospective, multi-centre validation, calibration assessment, subgroup fairness analysis, and integration with electronic health-record systems. The proposed framework may serve as a template for trustworthy AI systems in cardiology and other evidence-rich clinical domains.

Works Cited and Consulted

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In Proceedings of the International Conference on Learning Representations (ICLR).

Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., et al. (2023). Meditron-70B: Scaling medical pretraining for large language models. arXiv preprint arXiv:2311.16079.

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. https://doi.org/10.1136/bmj-2023-078378

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., et al. (2024). From local to global: A graph RAG approach to query-focused summarization. arXiv preprint arXiv:2404.16130.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745-e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 25(1), 65-69. https://doi.org/10.1038/s41591-018-0268-3

Khan, S. S., Matsushita, K., Sang, Y., Ballew, S. H., Grams, M. E., Surapaneni, A., et al. (2024). Development and validation of the American Heart Association PREVENT equations. Circulation, 149(6), 430-449. https://doi.org/10.1161/CIRCULATIONAHA.123.067626

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (NeurIPS), 33, 9459-9474.

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., & SPIRIT-AI/CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nature Medicine, 26(9), 1364-1374. https://doi.org/10.1038/s41591-020-1034-x

ISSN: 3107-5088, Vol. 2, Issue-3 (July-Sep, 2026), pp. 113-118, www.cognitivethinking.in

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS), 30, 4765-4774.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 23(6), bbac409. https://doi.org/10.1093/bib/bbac409

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., et al. (2023). Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617.

Topol, E. J. (2019). Deep medicine: How artificial intelligence can make healthcare human again. Basic Books.

World Health Organization. (2024). Cardiovascular diseases (CVDs) fact sheet. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. In Findings of the Association for Computational Linguistics (ACL 2024), 6233-6251.

Declaration by Author (s): *"I/We hereby declare that this manuscript is my/our original work, free from plagiarism, and that all sources and any use of Artificial Intelligence tools for content generation or editing have been fully disclosed and verified for accuracy.”* Ashish Mathur