

Cardia: A Smartphone-Only Framework Integrating Contact Photoplethysmography and Self-Reflective LangGraph Agent for Health Informatics

Vignesh Naik¹, Sharon Abreo², Renata Pinto³, Yuvraj Hinde⁴, Kedar Sawant⁵ Shrikrishna Narvekar⁶

^{1,2,3,4,5,6}Department of Computer Engineering, Agnel Institute of Engineering And Management

Abstract - This paper presents Cardia, a smartphone-based health agent designed to democratize access to personalized medical insights by removing the financial and physical barriers associated with traditional wearable hardware. The system uses an optimized contact photoplethysmography (cPPG) pipeline to extract biometric data, using standard smartphone cameras, along with a robust Multi-Agent System (MAS) built on LangGraph for safe, evidence-based medical reasoning. The methodology employs rigorous digital signal processing, including cubic spline interpolation, bandpass filtering, and ectopic rejection, to standardize measurements, which are interpreted by an agentic framework featuring cyclical, self-reflective safety loops. Empirical evaluation of the cPPG pipeline against clinical-grade pulse oximeters highlight high accuracy with a strong positive correlation and a Mean Absolute Error of 3.21 BPM. Furthermore, the agentic architecture successfully deals with generative AI risks, intercepting and correcting potential hallucinations or diagnostic overreach through critic node. Cardia establishes a scalable, safe, and privacy-focused ecosystem for proactive health management, providing a feasible pathway for the integration of intelligent agents in clinical and daily wellness environments.

Keywords: contact photoplethysmography, biometric monitoring, multi-agent systems, AI safety, healthcare, LangGraph.

1. INTRODUCTION

The real democratization of personalized healthcare depends on removing the physical and financial barriers to biometric data collection. In current time, recording physiological parameters depends heavily on wearable hardware, such as smartwatches, smart rings, and pulse oximeters. Although these devices provide valuable insights, their usually high costs indirectly create a divide, isolating 56% of populations who don't own wearables, from the growing preventative health system[1]. A vast number of people are left without the tools at their disposition to monitor their physiological well-being. The fast growth of Large Language Models (LLMs) has led to an increase in the number of users seeking conversational medical and wellness advice. The concept is deeply human, people want an empathetic, always available entity to understand their bodily state and stress levels. Deploying standard LLMs in user diagnostic or wellness is very dangerous. Traditional LLMs operate on single-pass architectures, they generate text sequentially and are subjected to "hallucinations", inventing facts, misinterpreting symptoms, or providing advice which does not consider the real-time physiological state of the

user. A standard chatbot cannot check its own medical advice before responding. When health is in question, confidence without base is a liability.

To mitigate these limitations of accessibility and safety, this paper introduces Cardia, a smartphone application that extracts real-time biometrics using only a smartphone camera and understands them through a self-correcting medical reasoning engine.

Our contribution is twofold, we put forth a highly optimized, cloud-processed contact Photoplethysmography (cPPG) pipeline. Cardia achieves comparable heart rate extraction to a clinical grade pulse oximeter, it turns any standard smartphone into a capable biometric scanner. Secondly, we introduce a Multi-Agent System (MAS) utilizing LangGraph to create an "Agentic RAG" (Retrieval-Augmented Generation) workflow. This system allows AI the autonomy to fetch peer-reviewed medical abstracts with enforcement of a cyclical, self-reflective safety loop. It's like having another medical personnel consulting on a patient's case, the AI must collaboratively verify its own outputs against strict safety constraints prior to user delivery[1].

2. LITERATURE REVIEW

Smartphone-based health monitoring and intelligent multi-agent systems have emerged as promising technologies for improving the accessibility and effectiveness of digital healthcare. The following review examines prior studies on PPG, HRV analysis, stress assessment, and LangGraph architectures to establish the research context for the proposed framework.

The study by Moya-Ramon et al. (2022) demonstrates that smartphone applications, specifically those utilizing chest straps (e.g., Elite HRV) or camera-based PPG (e.g., Welltory), provide a valid and reliable alternative to ECG for monitoring heart rate variability (HRV) in elite athletes. Their findings reveal almost perfect correlations ($r = 0.77$ to 0.94) and excellent intraclass reliability (ICC) for both standard 5-minute and ultra-short 1-minute recordings in supine and seated positions. This suggests that ultra-short measurements are sufficiently accurate for high-performance environments, allowing practitioners to efficiently monitor autonomic recovery and training adaptation without the need for clinical-grade equipment[2].

Bautista et al. (2023) conducted a systematic review and meta-analysis confirming that contactless photoplethysmography (cPPG) is a highly accurate, non-invasive alternative to ECG for heart rate monitoring in adults. Their analysis of 12 studies revealed a negligible mean bias of -0.13 bpm, demonstrating that camera-based sensors can reliably track vital signs without physical contact. While the

technology shows immense potential for remote triage and monitoring patients with skin sensitivities or delirium, the authors note that further validation is required for more complex metrics like SpO2 and HRV before widespread clinical adoption[3].

Singstad et al. (2021) evaluated the accuracy of finger photoplethysmography (PPG) against standard electrocardiography (ECG) for estimating heart rate variability (HRV) during rest, mild mental stress, and mild exercise. Analyzing synchronous five-minute recordings from 15 healthy participants, the authors demonstrated that PPG-derived HRV metrics specifically SDNN and RMSSD are highly accurate during stationary conditions but degrade significantly during exercise due to motion artifacts. Crucially, the study highlights that PPG-HRV reliability depends heavily on the selection of pulse fiducial points and robust outlier correction, with SDNN proving more resilient to noise than RMSSD[4].

Elgendi et al. (2019) provided a comprehensive review on the utility of photoplethysmography (PPG) as a non-invasive, digital health tool for assessing and monitoring hypertension. The authors highlighted that morphological features of the PPG waveform such as the digital volume pulse, crest time, and augmentation index correlate strongly with arterial stiffness and systemic blood pressure. While emphasizing PPG's potential for continuous, cuffless blood pressure monitoring via wearable devices, the study underscores critical challenges that must be addressed for clinical adoption, including the lack of standardized measurement protocols, vulnerability to motion artifacts, and the need for robust, diverse datasets to validate generalized machine learning models across different demographics[5].

Mathew Biju (2025) evaluated the architectural design, methodologies, and performance of multi-agent systems built using the LangGraph framework for large language model (LLM) applications. Through a case study implementing a centralized supervisor agent to orchestrate specialized sub-agents (e.g., database and FAQ retrievers), the author analyzed how LangGraph's graph-based structure optimizes information routing and complex task allocation. While demonstrating that the framework significantly enhances task execution accuracy and resource efficiency through structured state management, the study also highlights development bottlenecks, emphasizing that balancing agent orchestration requires careful benchmarking of task completion times and throughput to mitigate coordination overhead[6].

Jang and Li (2024) introduced AU-RAG, an agent-based framework designed to enhance Retrieval-Augmented Generation (RAG) by addressing the limitations of static retrieval processes in complex environments. By employing a multi-agent architecture that dynamically manages query planning, document retrieval, and answer synthesis, the authors demonstrated that Agent-based Universal RAG (AU-RAG) significantly improves the relevance and accuracy of generated responses across diverse data sources. Their evaluation indicates that this universal approach effectively mitigates common RAG failure points, such as irrelevant context and hallucination, though the authors note that the increased system complexity requires careful optimization of agent coordination to maintain low latency (Jang and Li, 2024)[7].

He et al. (2025) present a systematic literature review exploring the integration of Large Language Model (LLM)-based multi-agent systems (MAS) within software engineering, documenting the evolution from autonomous single-agent tools to sophisticated collaborative workflows. The authors establish a comprehensive taxonomy of MAS applications across the software development lifecycle including requirements engineering, code generation, and automated testing while characterizing prevalent architectural paradigms such as vertical and horizontal collaboration. The study synthesizes current research to outline a future development roadmap, emphasizing that while MAS enhances collective problem-solving capabilities, future advancement remains contingent upon overcoming significant hurdles in agent communication overhead, long-horizon reasoning consistency, and the establishment of standardized benchmarks for evaluating multi-agent performance in production environments[8].

Lei (2024) provides a comparative technical analysis of the LangGraph and LangChain frameworks, evaluating their respective efficacy in orchestrating autonomous AI agents. The author delineates the architectural shift from LangChain's linear, Directed Acyclic Graph (DAG) structures which often struggle with complex, iterative reasoning to LangGraph's stateful, cyclic graph paradigm. By formalizing agent interactions as nodes and edges with persistent state management, Lei demonstrates that LangGraph offers superior control over loops and conditional transitions, critical for non-linear task execution. The study concludes that while LangChain remains highly effective for rapid prototyping of straightforward chains, LangGraph is the more robust solution for developing high-order, multi-agent systems that require deterministic cyclical logic and sophisticated error-recovery mechanisms[9].

Spieser et al. (2026) systematically review the deployment of multi-agent systems (MAS) in biomedical domains, demonstrating how collaborative agent networks overcome the architectural limits of standalone LLMs to improve performance in oncology decision-making and clinical diagnostics. However, the authors formalize a critical "unreliability tax" within these workflows, characterized by a 15–50-fold increase in token consumption and a compounding risk of cascading hallucinations across agent pipelines. Ultimately, the study concludes that clinical adoption remains contingent upon transitioning from stochastic prompting to deterministic orchestration layers, supported by rigorous cost-utility frameworks to preserve clinical safety and expert oversight (Spieser et al., 2026)[10].

Duan and Wang (2024) explored the architectural synthesis of LangGraph and CrewAI to develop deterministic multi-agent applications for enterprise environments, specifically within telecommunication infrastructures. By embedding CrewAI's role-based collaborative agents within LangGraph's cyclic state-machine orchestration, the authors formalized a hybrid design paradigm that mitigates the unpredictable execution paths of autonomous agents. Implementation results in automatic log analysis and customer service pipelines demonstrate that this dual-layer framework optimizes state routing and memory persistence while enforcing deterministic guardrails. Ultimately, the study concludes that combining CrewAI's granular task execution with LangGraph's structured flow control significantly enhances system reliability and

minimizes cascading execution failures in production-grade workflows[11].

Zaragosa (2025) addresses the critical challenge of factual inaccuracies and "hallucinations" in Large Language Models (LLMs) by implementing a specialized Retrieval-Augmented Generation (RAG) framework for PubMed. The proposed architecture integrates ChromaDB for the semantic indexing of biomedical abstracts and Google's Gemini for text synthesis, thereby grounding the model's generative output in verifiable, real-time data. By transitioning the LLM from a static knowledge base to a dynamic reasoning engine, this approach significantly enhances the accuracy and reliability of information retrieval, offering a scalable and evidence-based solution for clinical decision support systems[12].

Huizinga et al. (2025) evaluate the efficacy of the Baevsky Stress Index (SI) as a non-invasive tool for measuring autonomic sympathetic tone and reactivity based on heart rate variability (HRV). Recognizing that sympathetic dysfunction is closely linked to chronic systemic and gastrointestinal motor disorders, the authors propose an optimized SI calculation methodology evaluated across 73 healthy subjects stratified into distinct age groups. Through Gaussian mixture modeling and kernel density analysis, the study establishes age-specific normative baselines for sympathetic activity at rest and during postural shifts (supine-to-standing), successfully identifying clinical thresholds for autonomic stress and dysfunction. Crucially, the authors validate this refined framework by demonstrating significant baseline variations in patients with chronic gastrointestinal motility disorders (such as chronic constipation), highlighting the index's utility as a diagnostic marker beyond traditional cardiology and providing a scalable tool for assessing autonomic health in clinical practice[13].

Di Credico et al. (2024) evaluate the autonomic and psychophysiological strains in motorsport by monitoring time-domain heart rate variability (HRV) metrics specifically rMSSD and SDNN in an elite rally crew during competition. Longitudinal tracking revealed a marked shift toward sympathetic dominance and elevated Baevsky Stress Indexes on race days, characterized by a constriction in Poincaré plot geometric distributions. Despite showing parallel physiological stress profiles, the driver exhibited greater subjective perception of fatigue than the co-driver, demonstrating that while both roles incur severe neurovegetative loads, daily HRV monitoring is essential to tailor individualized recovery and performance protocols in high-stress athletic environments[14].

3. METHODOLOGY

PART A: cPPG Scan Pipeline

To make monitoring of biometrics accessible on smartphones, Cardia implements a Thin-Client Architecture. The mobile application acts as a secure interface for collecting data, sending raw video data asynchronously to a Celery and FastAPI based backend where the heavy computation occurs.

1. Capturing the Scan and Video Handling:

The user experience begins with a physical scan. The user is prompted to place their index finger without heavy pressure, over the smartphone's rear camera lens and the adjacent LED

flash. The React Native frontend captures a 30-second video of the finger with its capillary bed.

Once the scan is complete, the .mp4 file is sent using an encrypted HTTP POST request to the FastAPI backend, where it is temporarily saved in the memory. Point to note here is that privacy is enforced at the hardware level to build absolute user trust. Once OpenCV processes through the video to extract the green-channel pixel density frame-by-frame, the temporary video file is permanently destroyed. No video of the user is ever stored or viewed.

2. DSP Processing Pipeline:

The signals which we extract from the video data are inherently noisy. It is polluted by micro-movements, breathing patterns, and variable hardware performance. To find the real cardiac rhythm, the data goes through a rigorous Digital Signal Processing (DSP) pipeline, as shown in Figure 1:

- **Interpolation:** Compared to a dedicated medical monitor, a user's smartphone under influence of multiple running apps will stutter and result in a signal where some frames are dropped and so on. To mitigate this we use Cubic Spline Interpolation method to mathematically standardize the time-domain of the signal to 30 frames per second (FPS).
- **Detrending:** Not all users will press the finger with the same pressure, and when that happens it results in changes in the baseline. To calculate the slow trend and mathematically subtract the unwanted changes we use the Savitzky-Golay filter, which gives us only the sharp micro-pulsations of the blood.
- **Bandpass Filtering:** To get only the human pulse avoiding any unnecessary influence, a 4th-order Butterworth bandpass filter is applied between 0.75 Hz and 3.0 Hz. This eliminates any frequencies outside the limits of the human resting heart (45 BPM to 180 BPM).
- **Ectopic Rejection:** An unexpected finger twitch can look like an artificial heartbeat to an algorithm. To avoid motion artifacts from ruining the variance mathematics, an ectopic rejection method separates the median RR interval (the time between beats) and dynamically removes any detected peak that shows deviation by more than 25% from the median.

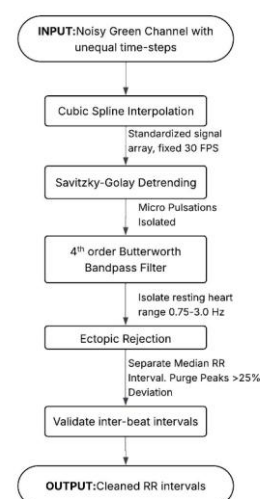


Fig -1: Digital Signal Processing Pipeline to extract RR intervals from the cPPG Scan

3. Metric Extraction and Output Handling:

Once the valid RR intervals (measured in milliseconds) are determined, they are translated into standard metrics:

a) Core Time-Domain Metrics:

The time divided array of inter-beat intervals, denoted as RR, lets us extract physiological baselines:

Heart Rate (HR): Calculated as the standard frequency:

$$HR_{bpm} = \frac{60,000}{\overline{RR}}$$

SDNN (Standard Deviation of NN intervals): Shows the overall flexibility and regulatory limits of the autonomic nervous system:

$$SDNN = \sqrt{\frac{1}{N} \sum_{i=1}^N (RR_i - \overline{RR})^2}$$

RMSSD (Root Mean Square of Successive Differences): The main time-domain measure used to determine high-frequency parasympathetic vagal tone:

$$RMSSD = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N-1} (RR_{i+1} - RR_i)^2}$$

b) Histogram-Based Stress Evaluation:

To calculate the stress, the app calculates the Baevsky Stress Index (SI), originally developed for aerospace medicine to monitor astronauts under extreme conditions. The RR intervals are divided into 50ms bins to determine the amplitude of the mode (AMo50), which represent the percentage of intervals matching the most frequent duration, and the variational sweep (MxDMn), which is the difference between the longest and shortest intervals:

$$SI = \frac{AMo_{50}}{2 \cdot Mo \cdot MxDMn}$$

Note: Mo and MxDMn are converted to seconds for this calculation.

c) Heuristic Meta-Scoring (0-100 Scales)

Raw physiological parameters, such as an RMSSD of 15ms or a Baevsky Index of 300, are usually opaque to the average user and difficult for Large Language Models to find context without resorting to hallucinations. To solve this, the raw metrics are algorithmically mapped onto 0-100 heuristic scales (Stress, Energy, Health, and Focus).

It is important to note that these meta-scores are not designed, nor claimed, to be clinical diagnostic metrics. Instead, they act as a translation layer, limiting high variance into percentiles to safely increase the AI's understanding.

- **Stress Score:**
The stress heuristic linearly normalizes the Baevsky Stress Index (SI). A healthy resting SI falls between

50 and 150, while values approaching 500 indicate high stress. The score limits at 100:

$$Stress = \max \left(0, \min \left(100, \frac{SI}{500} \cdot 100 \right) \right)$$

- **Health Score:**
The health score represents the overall flexibility of the autonomic nervous system by calculating the Standard Deviation of NN intervals (SDNN). Utilizing well proven norms where an SDNN of approximately 120ms represents absolute peak athletic condition, the array is scaled as follows:

$$Health = \max \left(0, \min \left(100, \frac{SDNN}{120} \cdot 100 \right) \right)$$

- **Energy Score:**
This metric is calculated from the natural logarithm of the RMSSD, which is the gold standard for tracking parasympathetic or vagal recovery. Since RMSSD is highly non-linear, applying a natural logarithm provides a more stable curve. The score is further penalized by elevated resting heart rates (HR > 90 BPM) to take into consideration cardiovascular fatigue:

$$Energy = \max \left(0, \min \left(100, \left(\frac{\ln(RMSSD)}{5.2} \cdot 100 \right) - \max(0, (HR - 90) \cdot 0.5) \right) \right)$$

- **Focus Score:**
The focus score is calculated after the Yerkes-Dodson Law (an inverted-U model of arousal). Peak cognitive focus occurs at an homeostatic baseline where the Baevsky SI rests near 100. The score reduces if the user is too lethargic (SI < 50) or is going through high autonomic tension (SI > 150):

$$Focus = \max \left(0, \min \left(100, 100 - \frac{|SI - 100|}{3} \right) \right)$$

PART B: Autonomous Agentic Architecture

The biometric data generated using cPPG Scan acts as a state context for the agent.

1. Chat Entry Modes (With and Without Context): There are two ways to start interacting with the agent. If a user enters the chat after a cPPG scan, the AgentState is given the qualitative translation of the biometrics (e.g., "highly elevated heart rate", "low vagal tone"). Another way users can enter is without a scan, making the agent rely only on textual input and historical data, along with user profile.

2. Structure of the Agent: The LangGraph orchestrates a Finite State Machine (FSM), which inherently implements a Self Reflecting Multi Agent RAG System as shown in the Figure 2, it includes several distinct nodes:

Safety Gate (MS Presidio): Checks the query first, anonymizes PII using Microsoft Presidio, and checks for crisis keywords (e.g., "heart attack"), leading directly to an emergency response if necessary.

Heuristics Engine: Translates raw quantitative biometric data into qualitative textual summary.

Supervisor Node (Agent 1): It acts as a router. It reads the query and biometrics and allocates the task to the appropriate specialist (Clinical, Diet, Lifestyle, or General).

Specialist Nodes (Agent 2): These are domain experts that utilize tools (PubMed API for clinical literature, Tavily API for walled-garden web search) to draft a proof-based response.

Critic Node (Agent 3): This node acts like a reviewer to ensure that the response is not hallucinated and is backed by evidence available.

Deanonymize Node: Restores PII mapping for user message.

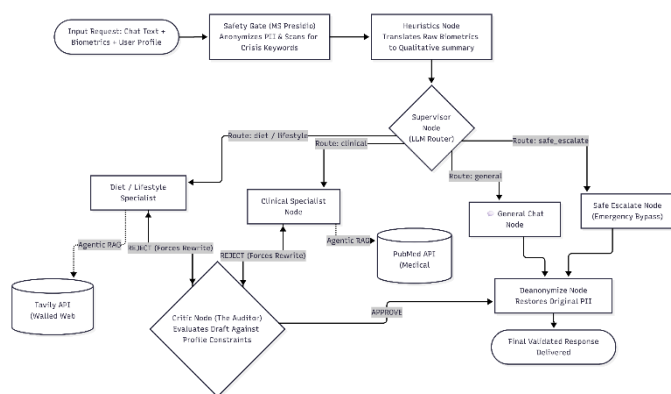


Fig -2: Self Reflecting Multi agent RAG Workflow utilized by Cardia

4. RESULTS AND EVALUATIONS

PART A: cPPG Scan Pipeline

To prove that the heart rate we read using the phone camera, using the digital signal processing technique mentioned in the Methodology section, which uses the contact photoplethysmography at the core to extract the heart rate using the change in the green channel of the video recording of the index finger, frame by frame, is equivalent to and in close range of the readings we get from a FDA approved pulse oximeter.

a) Ground Truth Testing

Mathematical accuracy of the scanning pipeline is very important, hence, empirical testing was organised against an FDA-approved clinical pulse oximeter across 120 independent trials (N=120). Table 1 describes the parameters collected during the ground truth testing process. Figure 3 shows the composition of both the genders in the collected dataset.

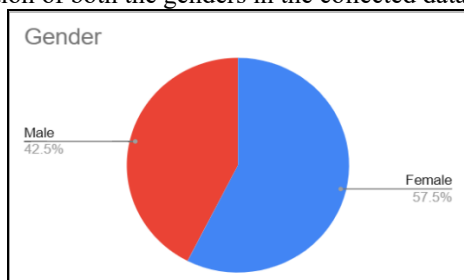


Fig -3: Gender distribution of the participants in the Comparison Dataset

Table -1: Parameters of the GTT

Parameter	Description
Participant_ID	Participant identifier
Age	Age of the participant
Gender	Male/Female
Oximeter_BPM	Average BPM recorded using the pulse oximeter
Cardia_HR_BPM	BPM estimated by the developed application
Difference_BPM	Difference between reference BPM and application BPM
Abs_Error	Absolute error between the two measurements

Correlation: Oximeter vs Cardia

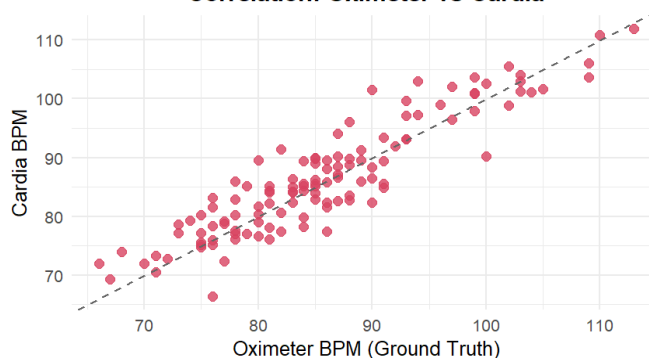


Fig -4: Correlation analysis of Cardia and pulse oximeter heart rate measurements.

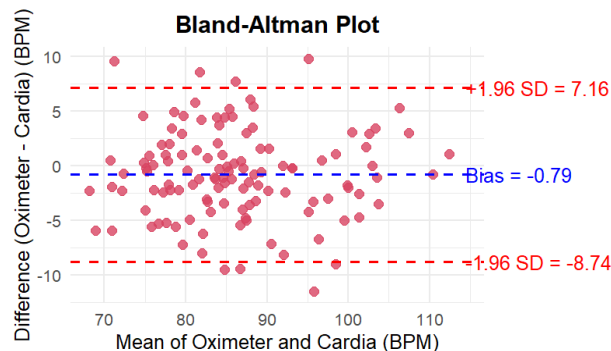


Fig -5: Bland-Altman plot comparing Cardia and pulse oximeter heart rate measurements.

The smartphone cPPG system built by us, demonstrated very high accuracy, achieving a Mean Absolute Error (MAE) of 3.21 BPM and Root Mean Square Error (RMSE) of 4.12 BPM factoring in the outliers. The Pearson correlation coefficient(r) was obtained to be 0.913 , which shows strong positive correlation. Figure 4 shows the same correlation in visual format. Figure 5 shows the Bland-Altman analysis to compare the measurements taken from the app and ones taken from the ground truth. It shows a range from -8 to + 7 bpm with mean bias of -0.79 bpm.

b) Statistical Equivalence

To test the viability of the cPPG extraction pipeline against the ground-truth pulse oximeter, a Two One-Sided Tests (TOST) for Equivalence was carried out on the dataset (N = 120). While

standard paired t-tests evaluate for statistical difference, which can become overly sensitive and flag clinically negligible variances in larger sample sizes, the TOST methodology explicitly tests for true statistical "agreement". The equivalence boundary was established at $\sigma = \pm 3.0$ BPM, strictly aligning with standard regulatory acceptance thresholds. The TOST evaluates the null hypothesis (H_0) that the absolute difference between the application's readings and the medical oximeter exceeds the acceptable ± 3.0 BPM margin (i.e., they are not equivalent). The testing yielded a near-zero mean measurement bias of -0.79 BPM. Both the upper and lower bound tests returned profoundly significant p-values ($p < 0.0001$). Because the maximum TOST p-value ($p < 0.0001$) is substantially lower than the standard 0.05 alpha level, we successfully reject the null hypothesis of non-equivalence. In summary, we have mathematically proven with greater than 99.9% confidence that the heart rate output generated by analyzing photometric color changes through a smartphone camera is statistically and clinically equivalent to the outputs of a dedicated clinical-grade pulse oximeter.

PART B: Autonomous Agentic Architecture

The evaluation of the LangGraph Agent was conducted using a set of 10 queries ranging from diet, lifestyle and medical domains, run for 3 different health profiles, you can refer to APPENDIX I for the 10 queries used for evaluations. The three profiles P1, P2 and P3 were as follows:

P1 : "Age:2002-05-14 | Gender: Male | Goal: Maintenance | Diet: Standard | Allergies/Aversions: None | Medical Conditions: None | Additional Info: None"

P2: "Age:2000-11-22 | Gender:Female | Goal:Muscle Gain | Diet: Vegetarian | Allergies/Aversions: Eggs, Peanuts, Tree Nuts | Medical Conditions: None | Additional Info: Highly sensitive to cross-contamination."

P3: "Age: 1985-04-10 | Gender: Male | Goal: Weight Loss | Diet: DASH Diet | Allergies/Aversions: None | Medical Conditions: Hypertension, Asthma | Additional Info: Takes blood pressure medication."

a)Execution Traces

The true power of LangGraph as the agent building framework is the advantage of using LangSmith, which allows us to track the full usage of the agent, including giving us a clear view of the agent traces, enhancing the value and use of explainable AI:

Medical Query Trace: A user, feeling exhausted, asks, "Why is my HRV so low today?" → Presidio clears PII → Supervisor identifies a medical query and routes to the Clinical Specialist → Specialist triggers the PubMed tool to fetch abstracts on "vagal tone and acute fatigue" → Draft generated → Critic reviews the draft, confirms peer-reviewed citations are present and that the tone is appropriately empathetic to the user's fatigue → Response delivered.

Lifestyle Query Trace: A stressed user asks, "What exercises can lower my stress index right now?" → Supervisor routes to the Lifestyle Specialist → Specialist queries Tavily for current, stress-reduction workouts → Draft generated → Critic rejects the draft because it robotically quoted the user's exact numerical stress index back to them (violating a core conversational tone constraint) → Graph routes backward to Specialist → Specialist rewrites the draft with a natural, human

tone → Rewrite approved by Critic → Response safely delivered.

b)Evaluation Metrics

During the above described evaluation test, the Self-Reflective Critic Node successfully intercepted and rejected 10% of initial LLM drafts. The primary reasons for rejection were diagnostic overreach, where the AI attempts to definitively diagnose a condition rather than suggest or hallucinate medical citations, where the AI invents a fake PubMed ID. The cyclical architecture successfully forces rewrites in every single instance, preventing these dangerous hallucinations from ever reaching the user. The average end-to-end latency for a full multi-agent cycle including tool retrieval and multiple safety reflection loops, was measured at ~ 10.80 seconds.

5. CONCLUSION

Cardia successfully shows the feasibility of an integrated, AI-powered healthcare assistant designed to bridge critical information gaps in accessible, preventive health management. By combining smartphone-based contact photoplethysmography (cPPG) biometric extraction with a robust Multi-Agent System (MAS), the application gives users actionable, evidence-based insights across physiological, nutritional, and psychological domains.

The system's architecture, highlights self-reflective safety loops and Retrieval-Augmented Generation (RAG), successfully mitigating the inherent risks of generative AI while ensuring data privacy. Empirical evaluation proved the reliability of the cPPG pipeline for the use case and the effectiveness of self reflecting agents in stopping propagation of risky hallucinated advice. Cardia establishes a scalable ecosystem that empowers users with personalized, safe, and proactive health support. Future iterations will focus on expanding the verifiable knowledge base and optimizing edge performance to further lower barriers to effective, intelligent healthcare, building the way for sustained, long-term well-being.

REFERENCES

- [1] S. Canali, V. Schiaffonati, and A. Aliverti, "Challenges and recommendations for wearable devices in digital health: Data quality, interoperability, health equity, fairness," *PLOS Digit. Health*, vol. 1, no. 10, Art. no. e0000104, Oct. 2022, doi: 10.1371/journal.pdig.0000104.
- [2] M. Moya-Ramon et al., "Validity and reliability of different smartphones applications to measure HRV during short and ultra-short measurements in elite athletes," *Comput. Methods Programs Biomed.*, vol. 217, 106696, 2022, doi: 10.1016/j.cmpb.2022.106696.
- [3] M. J. Bautista et al., "Clinical applications of contactless photoplethysmography for monitoring in adults: A systematic review and meta-analysis," *J. Clin. Transl. Sci.*, vol. 7, no. 1, e129, 2023, doi: 10.1017/cts.2023.547.
- [4] B. J. Singstad et al., "Estimation of Heart Rate Variability from Finger Photoplethysmography During Rest, Mild Exercise and Mild Mental Stress," *J. Electr. Bioimpedance*, vol. 12, no. 1, pp. 89–102, Dec. 2021, doi: 10.2478/joeb-2021-0012.
- [5] M. Elgendi et al., "The use of photoplethysmography for assessing hypertension," *NPJ Digit. Med.*, vol. 2, 60, Jun. 2019, doi: 10.1038/s41746-019-0136-7.
- [6] S. M. Biju, "Implementing Multi-agent Systems Using LangGraph: A Comprehensive Study," in *Recent Advancements in Artificial Intelligence (ICRAAI 2025)*, R. Nayak, N. Mittal, A.

Khunteta, and M. Kumar, Eds., *Lecture Notes in Networks and Systems*, vol. 1468. Singapore: Springer, 2026, doi: 10.1007/978-981-96-7760-3_19.

[7] J. Jang and W.-S. Li, "AU-RAG: Agent-based Universal Retrieval Augmented Generation," in *Proc. 2024 Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval Asia Pac. Region (SIGIR-AP)*, 2024, pp. 2–11, doi: 10.1145/3673791.3698416.

[8] J. He, C. Treude, and D. Lo, "LLM-Based Multi-Agent Systems for Software Engineering: Literature Review, Vision, and the Road Ahead," *ACM Trans. Softw. Eng. Methodol.*, vol. 34, no. 5, 124, May 2025, doi: 10.1145/3712003.

[9] Zhang Lei, "LangGraph vs LangChain: Which Framework Is Best for AI Agents?," *MetaVision Journal of Multidisciplinary Studies*, vol. 1, no. 4, pp. 16–26, 2024.

[10] J. Spieser et al., "A Review of Multi-Agent AI Systems for Biological and Clinical Data Analysis," *Methods Protoc.*, vol. 9, no. 2, 33, Feb. 2026, doi: 10.3390/mps9020033.

[11] Z. Duan and J. Wang, "Exploration of LLM Multi-Agent Application Implementation Based on LangGraph+CrewAI," 2024, arXiv preprint arXiv:2411.18241.

[12] A. R. A. Zaragosa, "Semantic Search and Generative AI for PubMed: A RAG Approach with ChromaDB and Gemini," *Int. J. Eng. Trends Technol.*, vol. 73, no. 10, pp. 161–173, 2025, doi: 10.14445/22315381/IJETT-V73I10P113.

[13] J. D. Huizinga et al., "Determining autonomic sympathetic tone and reactivity using Baevsky's stress index," *Am. J. Physiol. Regul. Integr. Comp. Physiol.*, vol. 328, no. 5, pp. R562–R577, 2025, doi: 10.1152/ajpregu.00243.2024.

[14] A. Di Credico et al., "Heart rate variability, recovery and stress analysis of an elite rally driver and co-driver during a competition period," *Sci. Prog.*, vol. 107, no. 1, 368504231223034, Jan.–Mar. 2024, doi: 10.1177/00368504231223034.