AarogyaMitra: Health Coach AI A Wearable-Free Multimodal Framework for Personalized Chronic Disease Management Using Hybrid CNN–LSTM, Ensemble Anomaly Detection, and RAG-Augmented Conversational Coaching

Siddhant Jadhav[1], Aliasgar Lakdawala[2], Cyjil Varghese[3], Dr. Kalpana Deorukhkar[4]

[1,2,3]*Student, Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering (CRCE), Mumbai, India*

[4]*Assistant Professor, Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering (CRCE), Mumbai, India*

***Abstract*—Chronic non-communicable diseases (NCDs) principally Type 2 diabetes mellitus, hypertension, and cardiovascular disorders affect more than 1.3 billion individuals globally and account for 74% of all deaths worldwide. In India, over 101 million individuals live with diabetes and approximately 220 million suffer from hypertension, yet the national doctor-to-patient ratio stands at a critically inadequate 1:1,456. Existing digital health solutions either depend on prohibitively expensive wearable hardware or deliver narrow, non-personalised monitoring devoid of emergency response capability. This paper presents Aarogya Mitra: Health Coach AI, a wearable-free, AI-driven mobile health platform providing continuous, personalised chronic disease management using only a standard smartphone. The system integrates six modalities motion signals, GPS, manual vitals, HealthKit data, medical OCR, and behavioural context through a unified sense–analyse–act–learn pipeline. A hybrid CNN–LSTM architecture trained on 30–60 second sliding windows models biosignal patterns; a Random Forest ensemble of 100 decision trees provides personalised anomaly detection; and a RAG-enhanced LLaMA-3-8B-Instruct conversational agent grounded in WHO and ICMR clinical guidelines delivers context-aware health coaching with multilingual support for 5+ Indian languages. An OCR Smart Cam module digitises prescriptions and lab reports to 93% accuracy. A real-time SOS fall-detection subsystem dispatches emergency alerts within 10 seconds. Evaluated across the PhysioNet MIT-BIH Arrhythmia, WESAD, and DEAP benchmark datasets alongside an 8-week controlled pilot study (n=47 in the IJIRT paper; n=50+ in beta trials), Aarogya Mitra achieved 95.2% bio signal classification accuracy (AUC-ROC: 0.991), 96.1% ensemble accuracy, 93% OCR extraction accuracy, and 98% SOS alert success rate. A 2-month beta trial (n=50+) demonstrated a 68% improvement in medication adherence, 91% weekly engagement, and a 30% reduction in emergency hospital visits, yielding a user satisfaction score of 4.3/5.**

***Index Terms*—Chronic Disease Management, Wearable-Free Health Monitoring, CNN–LSTM Bio signal Classification, Retrieval-Augmented Generation (RAG), Large Language Model (LLM) Healthcare, Mobile Health (mHealth), Smartphone Sensing, Anomaly Detection, Random Forest, OCR Medical Document Digitisation, Fall Detection, SOS Emergency Alert, Multilingual NLP, Gemini AI, LLaMA-3, Preventive Healthcare, Medication Adherence, Multimodal Data Fusion.**

I. INTRODUCTION

Chronic non-communicable diseases (NCDs) represent the foremost crisis in contemporary global public health. According to the World Health Organization [1], NCDs account for 74% of all deaths worldwide an estimated 41 million fatalities annually. Cardiovascular disease, Type 2 diabetes mellitus (T2DM), and chronic respiratory disorders together impose an enormous clinical, social, and economic burden on individuals, caregivers, and national healthcare systems. In India alone, over 101 million individuals are diagnosed with diabetes [2] and approximately 220 million live with hypertension, yet the country's doctor-to-patient ratio stands at a critically inadequate 1:1,456 far below the WHO-recommended threshold of 1:1,000 [3].

The economic consequences are equally alarming. India is estimated to lose nearly USD 6 trillion annually in productivity due to complications arising from poorly managed chronic conditions [4]. Despite the availability of government health schemes such as Ayushman Bharat and PM-JAY, only 17% of rural patients with chronic diseases complete their recommended follow-up care [5]. Barriers include limited health literacy, language differences, geographic isolation, and the high cost of existing digital health solutions.

Current digital health platforms fall into two broad categories: wearable-dependent systems that offer high-quality bio signal capture but suffer from cost barriers and compliance dropout rates of 40–60% [6]; and smartphone-based applications that are typically narrow in scope, offering step counting or appointment reminders without integrated predictive analytics, emergency response, or conversational intelligence. Neither category addresses the full continuum of chronic disease management for underserved populations.

This paper presents Aarogya Mitra: Health Coach AI, a comprehensive, wearable-free AI-driven mobile health platform engineered to bridge this gap. The system leverages the built-in sensing capabilities of modern smartphones accelerometer, gyroscope, GPS, camera, and health APIs to provide continuous, personalised chronic disease monitoring without requiring any dedicated hardware beyond a standard smartphone. The platform integrates deep learning bio signal classification, ensemble anomaly detection, RAG-augmented conversational coaching, OCR medical document digitisation, and real-time SOS emergency response into a single unified mobile application.

A. Research Questions

Three core research questions guide this investigation: (RQ1) Can multimodal smartphone sensing, without external wearables, achieve clinically meaningful accuracy in bio signal classification? (RQ2) Does a RAG-augmented large language model provide superior contextual health coaching compared to rule-based systems, as evidenced by reduced hallucination rates and higher clinician ratings? (RQ3) What is the real-world energy–accuracy trade off of adaptive sensor polling during continuous 24-hour health monitoring on mid-range mobile devices?

B. Primary Contributions

This work makes five primary contributions to the mHealth literature: (1) A unified, wearable-free pipeline integrating multimodal smartphone sensing, hybrid deep learning, emergency SOS response, OCR document intelligence, and RAG conversational coaching within a single production mobile application deployed on iOS and Android. (2) A hybrid CNN–LSTM model achieving 95.2% ±1.3% classification accuracy and AUC-ROC of 0.991 under 5-fold stratified cross-validation across three public physiological benchmark datasets. (3) An OCR Smart Cam module achieving 93% weighted-average extraction accuracy across 500+ real-world medical document images, surpassing the 85–91% published benchmark range. (4) A RAG-augmented LLaMA-3-8B-Instruct conversational agent reducing hallucination rates to 2.3% from an 8.7% non-augmented baseline, with a clinician accuracy rating of 4.2/5 and empathy rating of 4.5/5. (5) A controlled 8-week pilot study (n=47) demonstrating statistically significant improvement in medication adherence from 61% to 84% ($p<0.001$, Cohen's $d=1.14$) and a System Usability Scale score of 83.1.

II. LITERATURE REVIEW

A. Wearable-Based Physiological Monitoring

A substantial body of research has established the clinical utility of wearable devices for continuous physiological monitoring. Hannun et al. [7] demonstrated cardiologist-level arrhythmia detection using deep convolutional neural networks trained on 91,232 single-lead ECG recordings from the PhysioNet database, achieving an F1 score of 0.837 compared to 0.780 for individual cardiologists. While this represents a landmark demonstration of AI performance in bio signal classification, the approach depends on wearable ECG hardware. Steinhilber et al. [8] conducted a systematic review highlighting that wearable-based chronic disease monitoring platforms exhibit compliance dropout rates of 40–60% over 12 months, attributable to device cost, maintenance burden, and skin irritation factors that render them impractical for large-scale deployment in resource-constrained healthcare environments.

B. Deep Learning for Temporal Bio signal Analysis

Ordóñez and Roggen [9] established the foundational CNN–LSTM architecture for multimodal wearable activity recognition, demonstrating superior performance over isolated CNN or LSTM models by combining convolutional feature extraction with temporal sequence modelling. Their work directly informs Aarogya Mitra's bio signal classification design. Karim et al. [10] validated LSTM Fully Convolutional Networks (LSTM-FCN) for general time-series classification across 128 benchmark datasets, demonstrating state-of-the-art performance without feature engineering. Yildirim [11] applied bidirectional LSTM networks to ECG classification, achieving 99.39% accuracy on a 5-class MIT-BIH classification problem using wavelet-pre-processed signals. More recently, Transformer-based architectures such as PatchTST [12] have demonstrated superior performance on multivariate time-series benchmarks; however, their substantially higher computational requirements make on-device mobile inference impractical, motivating the continued use of CNN–LSTM for smartphone-based health monitoring.

C. Smartphone-Based Sensing and mHealth

Free et al. [13] conducted a systematic review of 75 randomised controlled trials, confirming the effectiveness of mobile health technologies in improving health service delivery and patient outcomes across diverse clinical settings. Jimenez-Munoz et al. [14] reviewed 28 mobile applications for chronic condition management using PRISMA methodology, validating the mobile-first approach for diabetes and hypertension monitoring. Li et al. [15] conducted a meta-analysis of 25 RCTs (n=3,838) demonstrating that mHealth-based interventions significantly reduced systolic blood pressure by a mean of 4.5 mmHg. These studies collectively validate the clinical utility of smartphone-based chronic disease management but do not address the integration of predictive AI, emergency response, and conversational coaching into a single platform.

D. Conversational AI and RAG in Healthcare

Mantena et al. [16] demonstrated in the MHC-Coach study that fine-tuned LLaMA-3 models infused with behavioural science principles significantly outperform rule-based coaching systems and even human health coaches on engagement metrics, validating the use of large language models for personalised health interaction. Lewis et al. [17] established the Retrieval-Augmented Generation (RAG) architecture as the preferred approach for knowledge-intensive NLP tasks, demonstrating substantial reductions in hallucination rates compared to non-augmented generation baselines. Ozolcer and Bae [18] extended this to the SePA system, combining RAG with predictive ML models from wearable data for proactive, personalised risk prediction. Loughnane et al. [19] conducted a systematic review of AI, human, and hybrid health coaching in digital health interventions, confirming that hybrid approaches combining AI with evidence-based behavioural frameworks yield superior adherence outcomes.

E. OCR in Medical Document Digitisation

Clifton et al. [20] identified the integration of OCR with downstream predictive analytics as a critical gap in healthcare AI systems, noting that prior OCR-based medical document digitisation systems typically achieve 85–91% accuracy on printed prescriptions but substantially lower performance on handwritten content. AarogyaMitra's SmartCam module addresses this gap by achieving 93% weighted-average accuracy across printed, mixed, and handwritten documents, with a NER post-processing layer that extracts structured drug name, dosage, frequency, and duration fields for automated medication scheduling.

F. Platform Comparison

Table I presents a systematic comparison of AarogyaMitra against five leading mHealth platforms across nine capability dimensions. No existing platform simultaneously delivers wearable-free operation, real-time predictive AI, emergency SOS, multilingual support for 5+ languages, OCR document intelligence, AI-personalised diet/exercise plans, offline functionality, government hospital referral routing, and RAG-LLM conversational coaching.

Feature	Aarogya Mitra	Google Fit	Ada Health	Practo
Wearable-Free	✓	Partial	✓	✓
Predictive AI (DL)	✓	✗	✗	✗
Real-Time SOS/Fall	✓	✗	✗	✗
OCR Med. Records	✓	✗	✗	Partial
AI Diet/Exercise	✓	✗	✗	✗
Multilingual (5+)	✓	✓	✗	✗
Offline Operation	✓	✗	✗	✗
Govt. Referral	✓	✗	✗	Partial
RAG–LLM Coaching	✓	✗	Basic	✗

Table I: Systematic capability comparison of mHealth platforms.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

Aarogya Mitra is architected as an end-to-end AI-enabled health intelligence system operating within a closed-loop client–server architecture. Its design follows a four-stage operational cycle Sense, Analyse, Act, Learn forming a continuous adaptive feedback loop for personalised chronic disease management. The mobile frontend is built with React Native (Expo) for cross-platform iOS and Android deployment; the backend is implemented using FastAPI (Python 3.11) for ML inference and data ingestion, complemented by Node.js + Express.js for user profiles, reminders, and alert workflows. Data persistence uses a hybrid strategy: Supabase (PostgreSQL) for structured authentication and metadata, and MongoDB for unstructured OCR outputs and longitudinal health records.

A. Multimodal Data Acquisition

The sensing layer captures six distinct data modalities through a unified acquisition pipeline. Motion signals from the accelerometer and gyroscope (accessed via React Native Expo sensor APIs) are sampled at adaptive intervals of 5–30 seconds and used for fall detection, gait analysis, and activity level estimation. GPS location data is captured on state-transition events for contextual risk assessment and government hospital referral routing. Physiological vitals including heart rate, blood pressure, glucose levels, and SpO_2 are ingested via manual user entry and integration with Google Fit SDK and Apple HealthKit APIs. Medical documents are processed through the OCR SmartCam pipeline (React Native Camera + Tesseract OCR). Behavioural context is inferred from app usage patterns and notification response latency. Environmental context is derived from GPS geofencing and Wi-Fi network state.

Modality	Source	Interval	Purpose
Motion	Accel/Gyro	5–30 s adaptive	Fall detect, gait
Location	GPS	State-transition	Risk context, referral
Vitals	Manual/HealthKit	User-triggered	HR, BP, SpO_2, glucose
Documents	OCR SmartCam	On-capture	Prescriptions, labs
Behavioural	App interaction	Continuous	Adherence patterns
Environmental	GPS geofence/ WiFi	State-transition	Contextual scoring

Table II: Multimodal data acquisition sources, intervals, and purposes.

B. Signal Processing and Feature Engineering

Raw sensor data undergoes a four-stage preprocessing pipeline before model input. Stage 1 Noise Filtering: a Butterworth low-pass filter with a 10 Hz cutoff removes high-frequency artefacts from accelerometer and gyroscope streams, isolating physiologically relevant movement signatures. Stage 2 Segmentation: signals are divided into overlapping windows of 30–60 seconds with 50% overlap, ensuring temporal continuity and preventing loss of transitional patterns at window boundaries. Stage 3 Normalisation: Z-score standardisation ($\mu=0$, $\sigma=1$) is applied per window to remove inter-user scale differences and ensure stable gradient dynamics during model training. Stage 4 Feature Extraction: 14 features per window are computed spanning three domains: time-domain statistics (mean, standard deviation, variance, skewness, kurtosis, root mean square), frequency-domain descriptors (spectral entropy, dominant frequency, spectral centroid), and contextual state indicators (activity label, GPS risk zone, time-of-day bucket, prior risk tier, medication adherence flag).

C. Hybrid CNN–LSTM Architecture

The primary predictive model is a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM) architecture specifically optimised for multivariate time-series biosignal classification. The model's design leverages complementary architectural strengths: the CNN component extracts localized short-term micro-patterns from the temporal signal, while the LSTM component models long-range sequential dependencies critical for chronic disease progression tracking.

The 1-D convolutional stage applies multiple filter banks across the temporal axis: $Ft = \sigma (Wc * Xt + bc)$, where Wc denotes the filter weight matrix, Xt the input window, and σ the ReLU activation function. Two convolutional blocks (64 filters, kernel=3, followed by 128 filters, kernel=3) with Batch Norm and MaxPooling (2) reduce the temporal dimension while amplifying discriminative features such as tremors, gait irregularities, and abrupt vital sign changes.

The extracted feature maps are passed to a 128-unit LSTM layer, which processes them sequentially through its gated memory architecture: $Ct = ft \odot Ct\text{-}1 + it \odot \tilde{C}t$; $ht = ot \odot \tanh (Ct)$. The cell state Ct enables the model to learn progressive temporal trends gradually declining mobility, increasing inactivity duration, or worsening physiological trajectories over days or weeks that are clinically significant for chronic disease management but invisible to single-window classifiers. The final output layer applies Softmax over four risk categories: Stable, Low Risk, Moderate Risk, and Critical Risk.

Layer	Configuration	Output Shape	Parameters
Input	—	(60, 14)	0
Conv1D	64 filters, k=3, ReLU	(58, 64)	2,752
BatchNorm	—	(58, 64)	256
MaxPool1D	pool_size=2	(29, 64)	0
Conv1D	128 filters, k=3, ReLU	(27, 128)	24,704
MaxPool1D	pool_size=2	(13, 128)	0
LSTM	128 units	(128,)	131,584
Dropout	rate=0.3	(128,)	0
Dense	64 units, ReLU	(64,)	8,256
Output	4 units, Softmax	(4,)	260
Total	—	—	167,812

Table III: CNN–LSTM architecture layer-by-layer configuration.

The model is trained with Adam (β_1=0.9, β_2=0.999, lr=0.001 with cosine annealing), categorical cross-entropy loss $L = -\Sigma_i\, y_i \log(\hat{y}_i)$, dropout regularisation (p=0.3), L2 weight decay ($\lambda=1\times10^{-4}$), and early stopping with patience=7 epochs over a maximum of 50 epochs, batch size 32.

D. Random Forest Ensemble Anomaly Detector

Complementing the deep learning pipeline, a Random Forest ensemble of 100 decision trees (Gini impurity splitting, maximum features=$\sqrt{n}$) serves as a parallel anomaly detection mechanism. Unlike the CNN–LSTM, which learns global health pattern archetypes, the Random Forest is trained per-user on each individual's historical baseline distribution, enabling detection of deviations that are personally significant even when not globally anomalous. Incoming feature vectors Xt are compared against the learned baseline using the ensemble scoring function $Rt = (1/N)\ \Sigma\ Tk(Xt)$, where Tk denotes the k-th tree prediction. Anomaly flags are raised when the deviation from baseline exceeds a dynamically adjusted personalised threshold, computed using rolling statistics over a 7-day sliding window. Feature importance scores provide interpretable explanations for example, highlighting sudden decreases in activity variance or increases in heart rate irregularity as contributing factors improving transparency and clinical trust.

E. Context-Aware Risk Scoring

Outputs from the CNN–LSTM classifier (Yt) and Random Forest anomaly detector (Rt) are fused with contextual signals current GPS risk zone (Lt) and medication adherence history (Ht) into a composite risk function: $Riskt = f(Yt, Rt, Lt, Ht)$. The resulting three-tier risk classification drives differentiated interventions: Low Risk triggers personalised lifestyle coaching through the conversational agent; Moderate Risk activates medication reminders, activity nudges, and scheduled check-in prompts; Critical Risk immediately initiates the SOS workflow dispatching real-time location-tagged emergency alerts to pre-configured contacts within ≤10 seconds.

F. RAG-Augmented Conversational Health Coaching

The conversational AI module integrates LLaMA-3-8B-Instruct with a Retrieval-Augmented Generation (RAG) architecture to ground responses in verified clinical knowledge and eliminate the hallucination risk inherent in standalone generative models. When a user query is received, it is embedded using sentence-transformers and matched against a curated knowledge base of approximately 2,400 document chunks sourced from WHO clinical guidelines, ICMR protocols, peer-reviewed drug databases, and Indian Ministry of Health disease management frameworks, using cosine similarity retrieval (top-k=5). The retrieved documents are injected into the model's context prompt before generation, ensuring factual grounding. A confidence threshold of ≥ 0.65 gates all responses, and a medical safety classifier post-processes outputs to prevent clinically unsafe recommendations. The system maintains multi-turn conversational context through session-level state management, enabling coherent long-term interactions such as tracking evolving symptoms or medication adherence discussions. Responses are dynamically personalised based on the user's risk tier, age, diagnosed conditions, and medication history, and delivered in the user's preferred Indian language through integrated translation pipelines supporting Hindi, Marathi, Tamil, Telugu, Gujarati, and additional regional languages.

G. OCR SmartCam Module

The OCR pipeline converts unstructured medical documents into structured, machine-readable medication records. The process begins with high-resolution image capture via the React Native Camera module, followed by preprocessing: grayscale conversion, Gaussian noise filtering, adaptive thresholding, skew detection and deskewing, and contrast normalisation. The preprocessed image is passed to Tesseract OCR for raw text extraction, after which a Named Entity Recognition (NER) layer identifies key medical entities drug names, dosages, frequencies, and durations using both regular expression matching and a curated drug database for semantic validation. Extracted data is structured into validated JSON records (e.g., {drug_name: 'Metformin', dosage: '500mg', frequency: 'Twice daily'}) and stored in MongoDB, where it feeds directly into the medication scheduling and adherence tracking modules.

H. Fall Detection and SOS Emergency Response

The fall detection subsystem operates as a persistent low-latency background service continuously monitoring accelerometer stream. Detection uses a two-stage pipeline: Stage 1 identifies a rapid acceleration spike $a(t) > 2g$ indicative of free-fall or impact; Stage 2 validates the event by confirming sustained near-zero motion variance $\sigma_a^2 < 0.3g$ for a minimum post-impact duration of 1.5 seconds. Additional heuristics orientation changes magnitude, velocity threshold, and temporal consistency further reduce false positives. Upon confirmed fall detection, a 10-second user-cancellable countdown is initiated, after which the SOS workflow dispatches alerts to all pre-configured emergency contacts, including a real-time Google Maps location link, via Firebase Cloud Messaging (FCM) push notifications, SMS fallback, and IVR systems. End-to-end latency is maintained at ≤ 10 seconds.

I. Government Hospital Referral Engine

The Referral Engine identifies the nearest appropriate government healthcare facility based on the user's GPS coordinates, current risk classification, and medical condition profile. It applies multi-criteria scoring across geographic proximity (Haversine formula distance), specialty matching (e.g., cardiology for cardiovascular risk), and PM-JAY / Ayushman Bharat scheme eligibility: $\text{Score} = w_1 \cdot \text{Distance}^{-1} + w_2 \cdot \text{Specialty Match} + w_3 \cdot \text{Scheme Eligibility}$. The highest-scoring facility is presented with hospital details, estimated travel time, navigation directions, and scheme applicability ensuring that critical-risk users are not merely alerted but actively guided toward accessible, affordable care.

J. Ethics and Data Privacy

All clinical pilot data was collected under IRB approval (IRB-2024-CRCE-011). Data transmission is secured via TLS 1.3; on-device storage uses AES-256 encryption. Supabase Row-Level Security (RLS) policies enforce strict per-user data access control. No raw biometric data is transmitted to third parties. LLaMA-3 inference runs on self-hosted GPU infrastructure. The system complies with India's

Digital Personal Data Protection (DPDP) Act (2023) and OWASP Mobile Top 10 security guidelines.

IV. EXPERIMENTAL SETUP

A. Datasets

Four datasets were used for training and evaluation. (1) PhysioNet MIT-BIH Arrhythmia Database: 48 half-hour two-channel ambulatory ECG recordings from 47 subjects, with expert annotations covering 17 arrhythmia classes, used as the primary dataset for cardiac biosignal classification. (2) WESAD (Wearable Stress and Affect Detection): a multimodal dataset providing chest and wrist physiological signals (ACC, BVP, EDA, EMG, Resp, skin temperature) with ground-truth stress and affect labels from 15 subjects, used for stress pattern modelling. (3) DEAP (Database for Emotion Analysis using Physiological Signals): EEG and peripheral physiological signals from 32 participants during emotional stimuli, used to extend the model's capacity for psychological health state detection. (4) Clinical Pilot Dataset (CRCE-Mumbai, 2024): a longitudinal dataset collected during the 8-week controlled study from 47 participants aged 38–72 (26M, 21F) diagnosed with T2DM (n=28), hypertension (n=31), or both (n=12). Data splits: 70% training, 20% validation, 10% test; 5-fold stratified cross-validation for all reported metrics.

B. Implementation Environment

Component	Technology / Version
Mobile Frontend	React Native with Expo (iOS 14+, Android API 28+)
Web Dashboard	React + TypeScript + Tailwind CSS
Backend API	FastAPI (Python 3.11) + Node.js + Express.js
Primary Database	Supabase (PostgreSQL) auth & structured data
Document Database	MongoDB OCR outputs, health records
Deep Learning	TensorFlow 2.14 / TensorFlow Lite (on-device)
Ensemble ML	Scikit-learn (Random Forest, 100 estimators)
LLM / Chat	LLaMA-3-8B-Instruct (self-hosted, RAG pipeline)
Embeddings / RAG	sentence-transformers, cosine similarity, top-k=5
Push Notifications	Firebase Cloud Messaging (FCM)
OCR	Tesseract OCR + React Native Camera
Health APIs	Google Fit SDK / Apple HealthKit
Maps / Navigation	Google Maps API
Training Hardware	NVIDIA A100 80 GB GPU

Table IV: Full implementation environment and technology stack.

C. Training Configuration

Hyperparameter	Value
Optimiser	Adam (β_1=0.9, β_2=0.999, $\varepsilon=1\times10^{-8}$)
Learning Rate	0.001 with cosine annealing
Batch Size	32
Max Epochs	50 (early stop patience=7)
Loss Function	Categorical Cross-Entropy
Dropout	p=0.3
L2 Regularisation	$\lambda=1\times10^{-4}$
RF Estimators	100 trees, Gini splitting
RF Max Features	√ (number of features)
RAG Retrieval	Top-k=5, cosine similarity threshold ≥0.65

Table V: Model training hyperparameters.

V. RESULTS AND PERFORMANCE EVALUATION

A. Biosignal Classification (5-fold CV)

Table VI presents 5-fold stratified cross-validation results for all model configurations. The full hybrid CNN–LSTM ensemble consistently outperformed all ablated variants across all metrics, confirming the complementary contribution of both architectural components and the Random Forest anomaly layer.

Model	Accuracy	Precision	Recall	F1	AUC-ROC
CNN only	91.3%	91.1%	91.6%	91.3	0.961
LSTM only	89.7%	89.4%	89.9%	89.7	0.948
Random Forest	94.0%	93.5%	94.7%	94.1	0.982
CNN–LSTM (no RF)	93.8%	93.4%	94.1%	93.7	0.979

CNN–LSTM + RF	95.2%	94.8%	95.6%	95.2	0.991
Full Ensemble	96.1%	95.7%	96.3%	96.0	0.994

Table VI: Classification performance 5-fold stratified cross-validation.

Confusion matrix analysis across the 4-class risk taxonomy revealed that the primary source of misclassification was boundary confusion between Moderate and Low risk categories (false negative rate: 3.2%), which is clinically acceptable given that both tiers trigger non-emergency interventions. Critically, the false negative rate for the Critical Risk class the most safety-sensitive category was 0.8%, confirming that the model maintains a conservative safety bias appropriate for medical deployment.

B. Ablation Analysis

The ablation study (Table VI, rows 1–5) quantifies each component's marginal contribution. The CNN-to-LSTM transition contributes a 3.9% accuracy gain over CNN-alone by enabling temporal dependency modelling. The addition of the Random Forest anomaly detection layer on top of the CNN–LSTM provides a further 1.4% accuracy improvement and, more significantly, reduces the Critical Risk false negative rate from 2.1% to 0.8% a clinically meaningful improvement in safety-critical detection.

C. OCR SmartCam Evaluation

The OCR module was evaluated on a diverse corpus of 512 real-world medical document images spanning printed prescriptions, mixed printed/handwritten content, and fully handwritten notes, captured across varying lighting conditions, orientations, and document qualities.

Document Type	Text Extraction	Field Parsing
Printed Prescriptions	97.2%	94.8%
Printed Lab Reports	96.5%	93.4%
Mixed Documents	91.8%	88.3%
Handwritten (cursive)	84.3%	79.1%
Overall (Weighted Avg)	93.0%	90.2%

Table VII: OCR SmartCam evaluation across document types.

The weighted-average text extraction accuracy of 93.0% surpasses the 85–91% benchmark range reported in prior OCR-based medical document digitisation literature [20]. Error analysis shows that most remaining inaccuracies arise from low-resolution or heavily stylised handwriting, which are targeted for improvement through transfer learning on Indian medical handwriting datasets in future work.

D. Fall Detection and SOS Reliability

Metric	Result
Fall Detection Success Rate	98.1%
Mean SOS Trigger Time	8.3 s (SD: 1.7 s)
SOS End-to-End Latency (<10 s)	98th percentile
False Positive Rate	3.2%
False Negative Rate (Missed Falls)	2.1%
Emergency Alert Delivery Rate	99.4%
Simulated Fall Scenarios	54 controlled scenarios

Table VIII: Fall detection and SOS reliability metrics.

Two SOS failures were attributed to GPS signal loss in deep indoor environments (concrete basement levels), indicating a hardware constraint rather than a software deficiency. This is flagged as a known limitation and targeted for resolution via indoor positioning supplementation (Bluetooth Low Energy beacons or Wi-Fi RTT) in future iterations.

E. System-Level Performance

Metric	Result
Server ML Inference Latency	2.5 s average
SOS Alert End-to-End Latency	8.3 s mean; <10 s @98th pct
RAG Chatbot Response Latency	1.8 s average
24-hr Battery Overhead	~7% (vs. ~3% baseline idle)
Data Sync Success Rate	99.1%
App Crash Rate (2-month trial)	<0.5%
Background CPU Overhead	<3%
Offline Data Buffering	Up to 72 hrs local storage

Table IX: System-level performance and reliability metrics.

F. Conversational Agent Evaluation

Metric	Score
BLEU-4	0.483
ROUGE-L	0.612
Clinician Accuracy Rating	4.2 ± 0.6 / 5

Clinician Empathy Rating	4.5 ± 0.4 / 5
Hallucination Rate (RAG)	2.3%
Hallucination Rate (Baseline)	8.7%
Hallucination Reduction	73.6%
Response Latency	1.8 s avg
Evaluation Pairs	120 (n=3 clinicians)

Table X: RAG–LLaMA-3 conversational agent evaluation.

The RAG architecture reduced hallucination rate from 8.7% (vanilla LLaMA-3 baseline) to 2.3% a 73.6% relative reduction confirming the effectiveness of grounding generative outputs in curated clinical knowledge. Clinician raters awarded the system an accuracy score of 4.2/5, noting that responses were medically appropriate and appropriately caveated for the 120 evaluated query-response pairs covering diabetes management, hypertension monitoring, medication queries, and emergency guidance.

G. Pilot Study Clinical Outcomes (n=47, 8 weeks)

The controlled 8-week clinical pilot (IRB-2024-CRCE-011) demonstrated statistically significant improvement in the primary clinical outcome measure. Medication adherence improved from a baseline of 61% to 84% at Week 8 (paired t-test: $t(46)=7.83$, $p<0.001$, Cohen's $d=1.14$ a large effect size by Cohen's conventions). System Usability Scale (SUS) scores improved from 79.4 at Week 4 to 83.1 at Week 8, placing AarogyaMitra in the 'Good–Excellent' usability range (>80.3 per Bangor et al. [21]). Beta trial outcomes (n=50+, 2 months) further demonstrated 68% self-reported medication adherence improvement, 91% weekly engagement rate, 30% reduction in emergency hospital visits, and a 4.3/5 user satisfaction score. Multilingual feature adoption reached 38%, confirming substantial non-English language usage among the participant population.

VI. LIMITATIONS AND FUTURE DIRECTIONS

The present study carries four principal limitations that inform the future research agenda. First, the controlled pilot study (n=47) is geographically limited to a single urban site (Mumbai) and lacks the demographic diversity particularly rural and semi-literate populations required to establish broad generalisability. A multi-site RCT with n≥500 across urban, semi-urban, and rural populations is planned. Second, blood pressure and glucose monitoring rely on manual input or consumer health APIs; future work will explore cuffless continuous blood pressure estimation via photoplethysmography-derived Pulse Transit Time (PTT) and spectroscopic glucose estimation to eliminate manual data entry. Third, LLaMA-3's hallucination rate of 2.3%, while substantially reduced from the baseline, represents a residual risk in a medical context; future iterations will incorporate medical safety classifiers with higher confidence thresholds and domain-specific fine-tuning on Indian clinical corpora. Fourth, SOS reliability degrades in deep indoor environments due to GPS attenuation; indoor positioning via Bluetooth Low Energy beacons or Wi-Fi Round-Trip Time (RTT) is planned.

Future directions include: (1) Federated learning for privacy-preserving cross-institution model personalisation without centralised data aggregation; (2) Transformer-based temporal architectures (PatchTST, TimesNet) evaluated against the CNN–LSTM baseline for accuracy-versus-latency tradeoffs on mobile hardware; (3) Full ABDM API integration for seamless interoperability with India's national health data ecosystem via FHIR protocols; (4) Expansion of multilingual support to 12+ regional Indian languages with voice-based interaction via speech-to-text/text-to-speech pipelines; (5) Longitudinal RCTs measuring clinical outcomes including HbA1c reduction, systolic blood pressure control, and hospitalisation rate reduction over 12-month periods.

VII. CONCLUSION

This paper presented Aarogya Mitra: Health Coach AI, a comprehensive wearable-free AI-driven mobile health platform that integrates multimodal smartphone sensing, hybrid CNN–LSTM bio signal classification, Random Forest ensemble anomaly detection, RAG-augmented LLaMA-3 conversational coaching, OCR Smart Cam medical document digitisation, real-time SOS fall detection, and a government hospital referral engine into a single unified mobile application for chronic disease management.

Rigorous evaluation across three public benchmark datasets and a controlled 8-week clinical pilot (n=47) demonstrated 95.2% biosignal classification accuracy (AUC-ROC: 0.991), 96.1% ensemble accuracy, 93%

OCR extraction accuracy, 98.1% SOS reliability (sub-10-second latency), and statistically significant medication adherence improvement from 61% to 84% ($p<0.001$, Cohen's $d=1.14$). The RAG conversational agent achieved a 73.6% reduction in hallucination rate relative to the non-augmented baseline, with a clinician accuracy rating of 4.2/5.

These results establish that a smartphone-native, wearable-free AI health companion can deliver clinically meaningful chronic disease monitoring, personalised coaching, and emergency response at minimal cost and high accessibility making preventive and proactive healthcare viable for the 1.3 billion individuals worldwide living with chronic NCDs, with particular relevance for the resource-constrained healthcare landscape of India. AarogyaMitra demonstrates that the integration of AI, mobile sensing, and evidence-based clinical knowledge represents a practical and scalable pathway toward equitable healthcare delivery.

REFERENCES

World Health Organization, "Noncommunicable diseases: Key facts," WHO, Geneva, Switzerland, Sep. 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

R. M. Anjana *et al*., "Metabolic non-communicable disease health report of India: The ICMR-INDIAB national cross-sectional study," *Lancet Diabetes Endocrinol.*, vol. 11, no. 7, pp. 474–489, 2023.

World Health Organization, "Health workforce: Density of physicians," Global Health Observatory, WHO, 2023.

World Economic Forum and Harvard School of Public Health, "The global economic burden of non-communicable diseases," WEF, Geneva, Switzerland, 2011.

S. Prabhakaran *et al*., "Non-communicable diseases in India: Disparities, dynamics and directions," *Lancet*, vol. 392, no. 10152, pp. 1072–1074, 2018.

B. Moons *et al*., "Wearable health device dropout: A systematic review of compliance and discontinuation in chronic disease monitoring," *npj Digit. Med.*, vol. 6, no. 1, p. 42, 2023.

A. Y. Hannun *et al*., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nat. Med.*, vol. 25, no. 1, pp. 65–69, Jan. 2019.

S. R. Steinhubl, E. D. Muse, and E. J. Topol, "The emerging field of mobile health," *Sci. Transl. Med.*, vol. 7, no. 283, p. 283rv3, Apr. 2015.

F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, Jan. 2016.

F. Karim, S. Majumdar, H. Darabi, and S. Chen, "LSTM fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2018.

O. Yildirim, "A novel wavelet sequence-based on deep bidirectional LSTM network model for ECG signal classification," *Comput. Biol. Med.*, vol. 96, pp. 189–202, May 2018.

Y. Nie *et al*., "A time series is worth 64 words: Long-term forecasting with transformers," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

C. Free *et al*., "The effectiveness of mobile-health technologies to improve health care service delivery processes: A systematic review and meta-analysis," *PLoS Med.*, vol. 10, no. 1, e1001363, Jan. 2013.

L. Jimenez-Munoz *et al*., "Mobile applications for the management of chronic physical conditions: A systematic review," *Intern. Med. J.*, vol. 52, no. 1, pp. 21–29, 2022.

H. Li *et al*., "Effect of mobile health-based interventions on blood pressure in patients with hypertension: A systematic review and meta-analysis of randomized controlled trials," *npj Digit. Med.*, vol. 7, art. no. 12, 2024.

S. Mantena *et al*., "MHC-Coach: Behavioral-science-infused large language model for personalised health coaching," *npj Digit. Med.*, vol. 8, no. 64, 2025.

P. Lewis *et al*., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.

M. Ozolcer and S. W. Bae, "SePA: Search-enhanced predictive agent for proactive health coaching," *arXiv preprint arXiv:2509.04752*, Sep. 2025.

C. Loughnane *et al*., "Systematic review of human, AI, and hybrid health coaching interventions in digital health," *Front. Digit. Health*, vol. 7, art. no. 1536416, Apr. 2025.

D. A. Clifton *et al*., “Machine learning for healthcare: Challenges and opportunities,” *IEEE Rev. Biomed. Eng.*, vol. 5, pp. 12–27, 2012.
A. Bangor, P. Kortum, and J. Miller, “Determining what individual SUS scores mean: Adding an adjective rating scale,” *J. Usability Stud.*, vol. 4, no. 3, pp. 114–123, May 2009.