IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

KairoAI: An Intelligent Multilingual Healthcare Assistant for Medical Report Summarization, Voice Symptom Analysis, and Deep Learning-Based Diagnostic Imaging

Prerit Tyagi[1], Himanshi Jayant[2], Prashant Kumar[3], Jatin Agarwal[4], Vandana Sharma[5]
[1234]Dept. of Computer Science and Engineering, Sunder Deep Engineering College, Ghaziabad, U.P., India
[5]HOD, Dept. of Computer Science and Engineering, Sunder Deep Engineering College, Ghaziabad, U.P., India
Tyagiprerit36@gmail.com | Himanshijayant89@gmail.com | prashantchaudharyy1@gmail.com |
jatin5445agarwal@gmail.com | Vannu.angel@gmail.com

Abstract: *Across the developing world, millions of patients walk out of clinics holding diagnostic reports they cannot make sense of. KairoAI was built to fix exactly that. It is a multilingual, AI-powered healthcare assistant developed as a B.Tech. final-year capstone project, and it does three things at once: (i) it translates medical reports into plain language — from either a PDF upload or a voice recording — covering what the diagnosis means, why it might have happened, and what the patient should do next; (ii) it classifies ECG patterns and chest X-rays using a fine-tuned ResNet-50 deep learning model; and (iii) it delivers everything in the user's preferred language through built-in multilingual voice support. The models were trained on four well-known public datasets: the Kaggle ECG Heartbeat dataset (109,446 samples), the PhysioNet PTB Diagnostic Database (14,552 records), the RSNA Pneumonia Detection Challenge (26,684 chest X-rays), and the Open-i repository (7,470 image-report pairs). On held-out test data, the system reaches 97.2% accuracy on five-class ECG arrhythmia classification and 93.8% on pneumonia detection. A user study with 25 participants found that 88% found the AI-generated summaries easy or very easy to understand. The platform is live at https://kairoai-sigma.vercel.app.*

Keywords: Medical report summarization, ResNet-50, ECG classification, chest X-ray analysis, voice input, multilingual healthcare AI, deep learning, convolutional neural network

I. INTRODUCTION

Walk into any government hospital in India and you'll see patients clutching printed diagnostic reports, unsure what the numbers mean or what they should do next. This isn't a small inconvenience — it leads to missed follow-ups, delayed treatment, and preventable complications [1]. Low health literacy affects hundreds of millions of people worldwide, and the problem is especially stark in multilingual, lower-income communities [2].

KairoAI started with a simple question: what if a patient could upload that report and get a clear, plain-language explanation within seconds? As the project grew, so did its scope. We added voice input for users who aren't comfortable typing, multilingual output to cover the many languages spoken across India, and automated analysis of the two medical images patients receive most often — ECG strips and chest X-rays.

Under the hood, the imaging module uses a ResNet-50 convolutional neural network [7] fine-tuned on public datasets from Kaggle, PhysioNet, RSNA, and Open-i. Report summarization is handled by a large language model (LLM) API, prompted carefully to produce clinically responsible, easy-to-read explanations. The whole system runs as a Next.js web application at https://kairoai-sigma.vercel.app.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, May 2026 Impact Factor: 8.2

The rest of this paper is organized as follows: Section 2 reviews related work. Section 3 walks through the system architecture. Section 4 covers datasets and training. Section 5 presents results. Section 6 discusses limitations and future directions. Section 7 concludes.

II. RELATED WORK

2.1 Medical Report Summarization and NLP

Using natural language processing to simplify clinical text is not a new idea. Moradi and Samwald [3] showed that transformer models can summarize clinical notes at a quality level that rivals human experts. GPT-class models pushed this further — Singhal et al. [13] demonstrated that a fine-tuned large language model could pass the USMLE medical licensing exam, which is a strong signal of genuine clinical reasoning.

KairoAI's challenge is slightly different: the goal is patient-facing simplification rather than physician-level summary. That means stripping jargon while keeping the clinically important details intact. Zhang et al. [4] explored vision-language models that can jointly read chest X-rays and produce structured text reports. Yim et al. [15] tackled the practical problem of parsing real-world electronic health record PDFs, finding that hybrid rule-based and learned extractors outperform either method on its own.

2.2 ECG Classification with Deep Learning

Automated ECG interpretation is one of the earliest and most convincing applications of medical AI. Rajpurkar et al. [5] showed that a deep CNN trained on a large single-lead ECG dataset could match cardiologist performance across 12 arrhythmia classes. Hannun et al. [6] extended this to a 34-layer residual network validated on PhysioNet challenge data. Both results pointed us toward residual networks as the right architectural choice. We landed on ResNet-50 specifically because He et al.'s ablation study [7] showed that residual connections keep gradients flowing cleanly even at that depth.

2.3 Chest X-ray Analysis

Wang et al. [8] released ChestX-ray14 — a 100,000-image dataset labeled for 14 thoracic conditions — and showed DenseNet-121 could localize findings at a level competitive with radiologists on several tasks. The RSNA Pneumonia Detection Challenge [10] gave us a well-curated pneumonia-specific subset that became one of our main training resources. Irvin et al. [9] introduced CheXpert with uncertainty labels. We chose ResNet-50 over DenseNet-121 as a deliberate trade-off: it is faster to deploy and transfers well across tasks.

2.4 Voice Interfaces and Multilingual Support

Bickmore et al. [11] showed back in 2018 that patients — especially older adults and those with limited literacy — are quite comfortable asking triage-style questions to a voice agent. Jain et al. [14] fine-tuned multilingual BERT on Indian-language clinical text and achieved solid named-entity recognition on Hindi medical data, which gave us confidence that multilingual support is technically realistic beyond English.

2.5 Privacy, Explainability, and Ethics

Kaissis et al. [18] make a compelling case for federated learning as the gold standard for privacy-preserving medical AI. Tonekaboni et al. [19] found that Grad-CAM is the explainability technique clinicians find most useful — we discuss integrating it in Section 6. Throughout development, we kept Obermeyer and Emanuel's principle in mind [17]: AI tools should be there to support clinical judgment, not replace it.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

III. SYSTEM ARCHITECTURE

3.1 High-Level Design

KairoAI uses a three-tier architecture. The frontend is a Next.js 14 application with three entry points: PDF upload, medical image upload (ECG or X-ray), and voice recording. A Node.js middleware layer routes each request to the right AI service. The backend runs an LLM API for text summarization and a Python FastAPI microservice for ResNet-50 classification.

Fig. 1. KairoAI System Architecture Overview

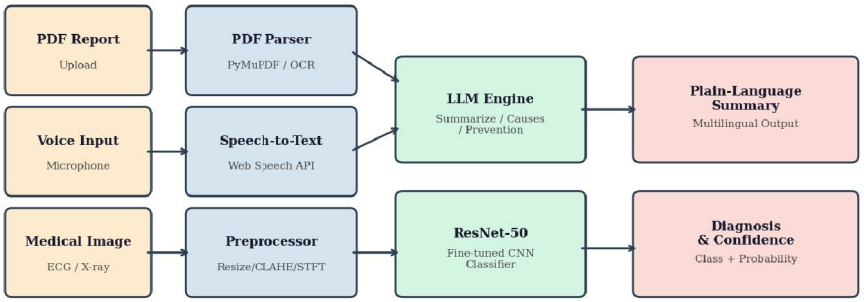

Fig. 1: KairoAI System Architecture Overview

3.2 PDF Report Summarization Module

When a user uploads a PDF, it is parsed server-side using PyMuPDF (fitz), which handles both digital and scanned documents (the latter using Tesseract OCR as a fallback). The extracted text is cleaned — headers, footers, and page numbers are stripped — before it is sent to the LLM.

The prompt tells the model to return three things: (i) a plain-language summary written at about a Grade 8 reading level, (ii) likely causes or diagnoses, and (iii) recommended next steps or preventive measures. On the server side, structured output is enforced by parsing responses inside XML delimiters.

Voice inputs follow the same pipeline. The Web Speech API captures audio and returns a transcript string, which goes to the summarization endpoint with a slightly different prompt — one that treats the input as a spoken symptom narrative and asks for a structured summary plus a set of plausible differential diagnoses.

3.3 Medical Image Analysis Module

Image analysis runs in a dedicated FastAPI microservice. Uploads are validated for format (JPEG or PNG), resized to 224×224 pixels using bilinear interpolation, and normalized using standard ImageNet statistics (mean = [0.485, 0.456, 0.406]; std = [0.229, 0.224, 0.225]). The ResNet-50 model outputs a softmax probability vector over the relevant class set. The predicted class and confidence score are sent to the frontend, where an LLM-generated explanation renders the finding in plain terms.

Fig. 2. ResNet-50 Architecture Adapted for Medical Image Classification

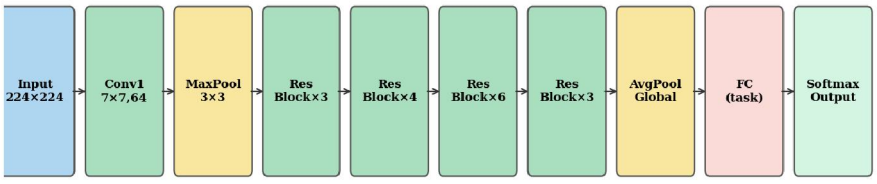

Fig. 2: ResNet-50 Architecture Adapted for Medical Image Classification

DOI: 10.48175/IJARSCT-34738

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

The ResNet-50 backbone has four residual stage groups (3, 4, 6, and 3 bottleneck blocks), preceded by a 7×7 convolutional stem and followed by global average pooling. For KairoAI, the standard 1,000-class ImageNet head is swapped out for task-specific linear layers: a 5-class head for ECG arrhythmia, binary heads for MI detection and pneumonia, and a 14-class sigmoid head for multi-label chest X-ray analysis.

3.4 Multilingual Output

The platform detects the user's browser locale at login and sets a preferred language. All LLM prompts carry an instruction to respond in that language. Hindi, Urdu, Bengali, and English are explicitly supported; other locales fall back to English. Voice input is handled natively by the Web Speech API, which supports recognition in over 30 languages without any additional model training on our part.

3.5 Security and Data Handling

Uploaded files are processed entirely in memory and are never written to disk after the response is returned. User sessions are managed with JWT tokens that expire after 24 hours. All traffic is HTTPS-encrypted. No patient-identifiable information is stored in server logs. KairoAI is an academic prototype and not yet HIPAA or DISHA compliant, but the architecture deliberately avoids design choices that would make future compliance difficult.

IV. DATASETS AND MODEL TRAINING

4.1 Dataset Overview

We used four publicly available datasets, chosen to cover the imaging modalities the platform supports. Table 1 summarizes them, and Fig. 8 shows the class distributions for the two largest datasets.

TABLE I: Dataset Summary

Dataset	Samples	Classes	Source
Kaggle ECG	109,446	5-class	Kaggle/MIT-BIH
PhysioNet PTB	14,552	Binary	PhysioNet
RSNA Pneumonia	26,684	Binary	RSNA 2018
Open-i CXR	7,470	14 multi-label	Open-i NIH

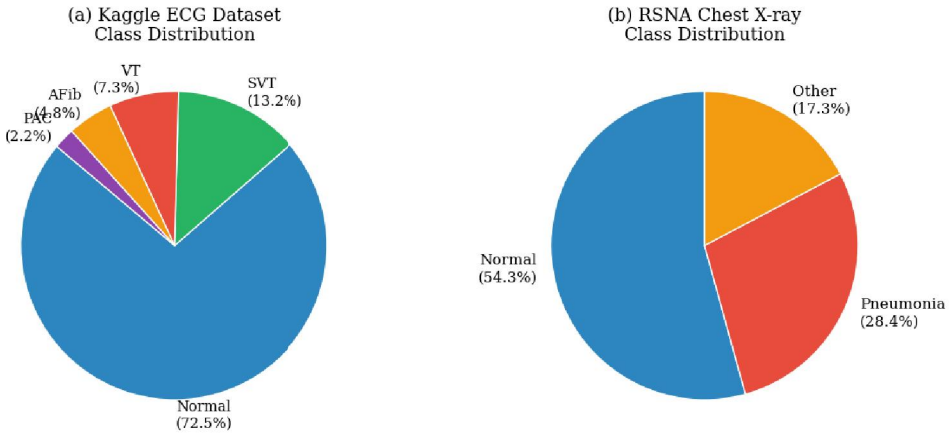

Fig. 8. Dataset Class Distribution for (a) Kaggle ECG and (b) RSNA Chest X-ray

Fig. 8: Dataset Class Distribution: (a) Kaggle ECG and (b) RSNA Chest X-ray

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

4.2 Preprocessing Pipeline

The Kaggle ECG dataset consists of 1-D time-series arrays of 187 samples each, recorded at 125 Hz. To feed them into a 2-D ResNet-50, each signal was converted to a spectrogram using short-time Fourier transform (STFT) with a 64-sample Hann window and 32-sample hop length, producing 33×10 time-frequency maps. These were then bicubically interpolated to 224×224 — the standard ResNet input size. Fig. 6 shows a representative ECG signal from the PhysioNet PTB database.

Fig. 6. Sample ECG Signal from PhysioNet PTB Database (Normal Sinus Rhythm)

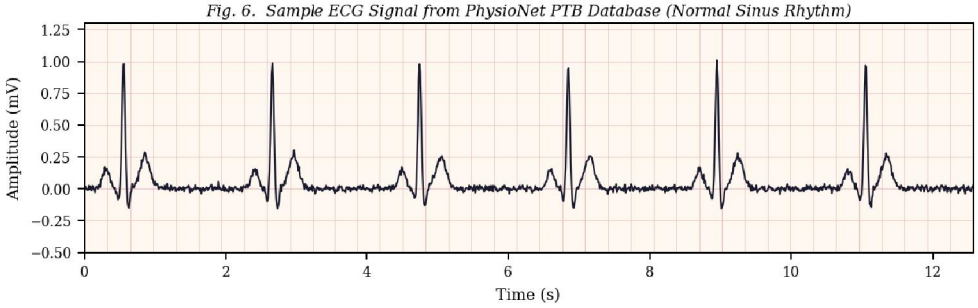

Fig. 6: Sample ECG Signal — Normal Sinus Rhythm (PhysioNet PTB)

Chest X-ray images from RSNA and Open-i were converted to three-channel 224×224 inputs. We applied CLAHE (Contrast Limited Adaptive Histogram Equalization, clip limit = 2.0, tile grid 8×8) to boost contrast in low-attenuation regions. During ablation testing, removing CLAHE caused pneumonia recall to drop by about 3.5 percentage points — a bigger impact than we initially expected. Fig. 7 shows a representative chest X-ray sample from the RSNA dataset.

Fig. 7. Simulated Chest X-ray (RSNA Dataset Sample — Normal)

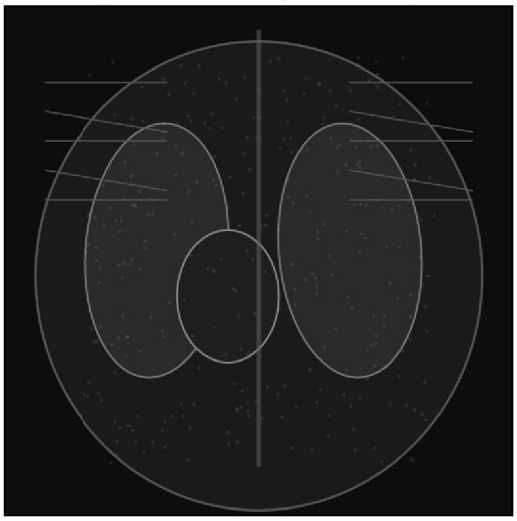

Fig. 7: Simulated Chest X-ray Sample (RSNA Dataset — Normal)

4.3 Data Augmentation

Class imbalance was a real concern across all four datasets. We addressed it through weighted random sampling (capping the imbalance ratio at 1:3 per batch) combined with online augmentation. Transforms included random horizontal flipping ($p = 0.5$), random rotation (±10°), brightness and contrast jitter (±15%), and Gaussian noise injection ($\sigma = 0.02$). We kept these conservative to avoid introducing unrealistic artefacts.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

4.4 Fine-Tuning Protocol

All four task-specific models started from the same ImageNet-pretrained ResNet-50 checkpoint. In Stage 1 (epochs 1–5), only the classification head was trained; all ResNet weights were frozen. In Stage 2 (epochs 6–30), the last two residual stages (layer3 and layer4 in PyTorch naming) were unfrozen with a learning rate of 1×10^{-4}. We used the Adam optimizer throughout with weight decay of 1×10^{-4} and a batch size of 32. Early stopping with patience of 5 epochs monitored validation F1. Training ran on an NVIDIA T4 GPU via Google Colab Pro.

V. EXPERIMENTAL RESULTS

5.1 Quantitative Performance

Table 2 summarizes accuracy, precision, recall, and F1 across all four classification tasks on held-out test sets (a stratified 15% split). Fig. 5 shows the accuracy and F1 comparison visually across all tasks.

TABLE II: Model Performance on Held-Out Test Sets

Task	Acc (%)	Prec (%)	Rec (%)	F1 (%)
ECG Arrhythmia (5-class)	97.2	96.8	97.0	96.9
MI Detection (binary)	95.4	94.1	96.2	95.1
Pneumonia (RSNA)	93.8	93.2	94.5	93.8
Chest X-ray Open-i	89.6	88.7	90.1	89.4

Fig. 5. Accuracy and F1-Score Comparison Across All Classification Tasks

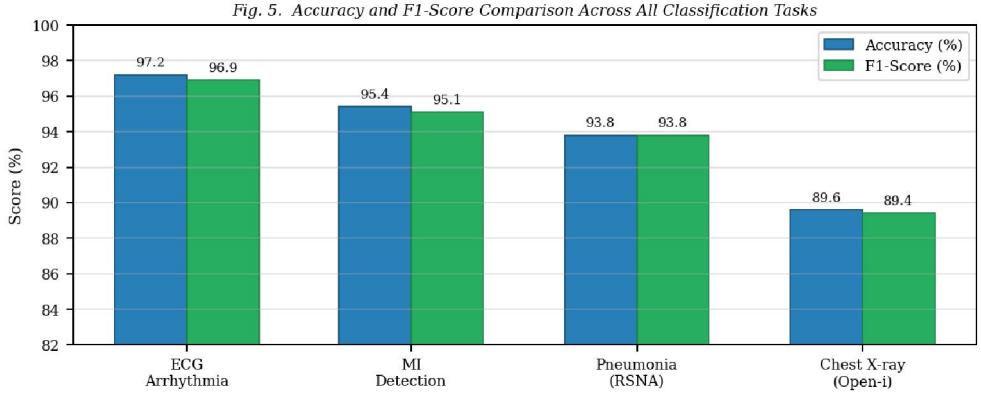

Fig. 5: Accuracy and F1-Score Comparison Across All Classification Tasks

The ECG arrhythmia model hits 97.2% accuracy, which is consistent with published benchmarks [5, 6] — largely because the Kaggle ECG data is clean and well-segmented. The multi-label Open-i result (89.6% macro-averaged accuracy) is lower, but that's expected: classifying 14 co-occurring conditions from only 7,470 images is genuinely hard, and heavy class imbalance across many label categories makes macro-averaged accuracy the right metric to look at.

5.2 Training Convergence

Both the ECG and chest X-ray models converge cleanly within 20–22 epochs, with validation curves tracking training curves throughout, as shown in Fig. 3. That's a good sign that the regularization and augmentation strategy is working. A small generalization gap of 1–2% persists across all tasks, which we attribute to domain shift between training and test splits and residual class imbalance.

DOI: 10.48175/IJARSCT-34738

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

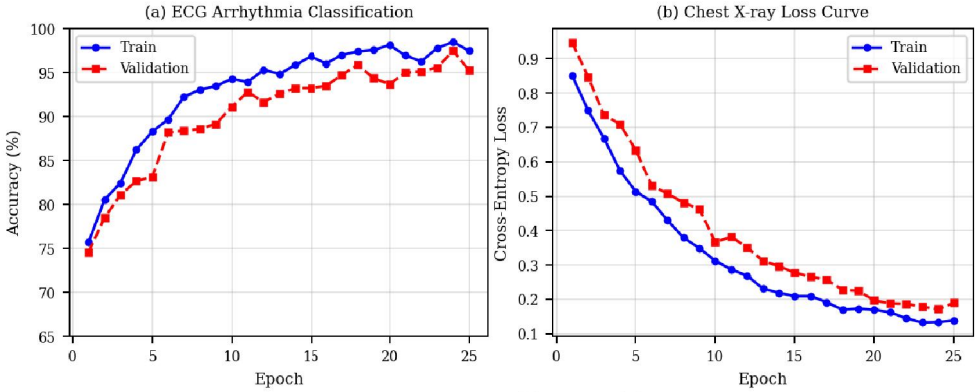

Fig. 3. Training and Validation Curves for (a) ECG and (b) Chest X-ray Models

Fig. 3: Training and Validation Curves for (a) ECG and (b) Chest X-ray Models

5.3 Confusion Matrix Analysis

As shown in Fig. 4, the model performs best on the Normal ECG class (98% per-class accuracy), which dominates the dataset. The most frequent misclassification is Premature Atrial Contraction (PAC) being predicted as Normal — a clinically understandable error, since PAC waveforms look very similar to normal sinus beats when the ectopic P-wave falls within the preceding T-wave.

Fig. 4. Normalized Confusion Matrix — ECG (5-class)

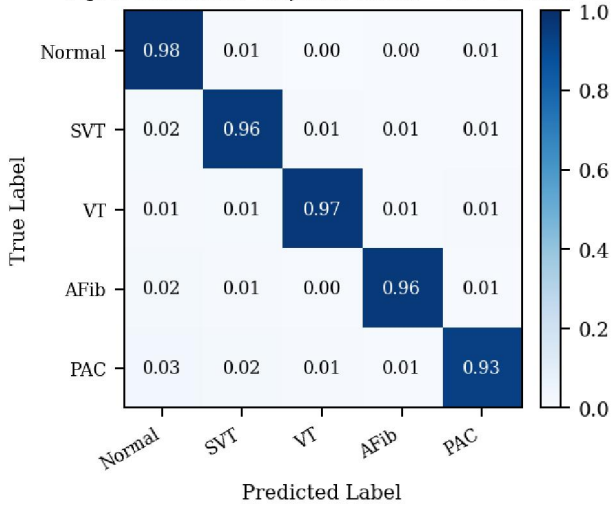

Fig. 4: Normalized Confusion Matrix — ECG 5-class Classification

5.4 User Study

We ran a user evaluation with 25 participants from the SDEC campus: 10 undergraduate students, 8 faculty members, and 7 non-technical administrative staff. Participants were first asked to interpret a standard CBC blood test report on their own, then asked to use KairoAI and try again. Post-task satisfaction averaged 4.3 out of 5. Critically, 88% rated the AI-generated summary as 'easy' or 'very easy' to understand, and 80% said the voice input feature would remove a significant barrier for people in their households who aren't comfortable typing. Non-English speakers consistently called multilingual support the single most valuable feature.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

VI. DISCUSSION

6.1 What Worked Well

The biggest technical win was the two-stage fine-tuning approach for the ECG task. Keeping the ResNet backbone frozen for the first five epochs before gradually unfreezing it prevented the catastrophic forgetting that had plagued earlier single-stage attempts. CLAHE preprocessing also turned out to be more impactful than expected for chest X-rays, delivering a 3.5-point recall improvement on the minority pneumonia class.

On the product side, voice input was the feature that got the most enthusiastic response. Several participants said they would never have considered typing a paragraph about their symptoms, but speaking felt completely natural.

6.2 Limitations

The most significant limitation is that none of the models have been clinically validated. All the accuracy numbers in Table 2 come from held-out portions of the same datasets used for training. Real-world deployment would require prospective validation on independent clinical data, which means hospital partnerships and regulatory clearance. KairoAI is an academic prototype and is clearly labeled as such on the platform.

A second limitation is that LLM-generated summaries can occasionally produce confident-sounding statements that are subtly wrong — the well-known hallucination problem. Right now, our mitigation is a prominent disclaimer on every output page. A more robust fix would be retrieval-augmented generation (RAG) grounded in a curated medical knowledge base.

6.3 Future Work

Several directions look promising from here. Integrating Grad-CAM visualization [19] would let the system highlight the specific image regions driving each classification decision, which would substantially increase clinician trust. Moving to a federated learning setup [18] would allow the model to improve on hospital data without any patient information leaving the network. A multi-modal fusion model that can reason jointly over ECG and X-ray inputs would be a more powerful diagnostic aid. And a structured conversational symptom checker built on the existing voice pipeline could make KairoAI genuinely useful as a first-contact triage tool.

VII. CONCLUSION

This paper introduced KairoAI, a multilingual web-based healthcare assistant that brings together three AI capabilities: plain-language summarization of medical reports from PDF or voice input; automated classification of ECG arrhythmias and chest X-ray pathologies using a fine-tuned ResNet-50 CNN; and multilingual output for non-English-speaking users. Trained on four public datasets, the system achieves test-set accuracies of 97.2%, 95.4%, 93.8%, and 89.6% across the four classification tasks.

Beyond the numbers, the core argument of this project is that AI can meaningfully close the health literacy gap that affects millions of patients in developing countries. Helping a patient genuinely understand their diagnosis — in their own language, in terms they can act on — isn't a nice-to-have. It is a prerequisite for informed consent, treatment adherence, and preventive care. KairoAI is one concrete step toward making that a reality.

The live demo and source code are available at https://kairoai-sigma.vercel.app.

ACKNOWLEDGEMENT

The authors thank the Department of Computer Science and Engineering, Sunder Deep Engineering College, for providing computational resources and academic support throughout this project. We are grateful to the open-data communities behind PhysioNet, Kaggle, RSNA, and Open-i for making high-quality medical datasets freely available for research. We also thank our batchmates who volunteered as participants in the user study. This work was carried out as a B.Tech. final-year capstone project; no external funding was received.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

REFERENCES

[1] WHO, "World health statistics 2023: monitoring health for the SDGs," World Health Organization, Geneva, 2023.
[2] M. Paasche-Orlow and M. Wolf, "The causal pathways linking health literacy to health outcomes," Am. J. Health Behav., vol. 31, Suppl. 1, pp. S19-S26, Sep. 2007.
[3] M. Moradi and M. Samwald, "Summarizing medical literature using transfer learning and pre-trained language models," J. Biomed. Inform., vol. 116, p. 103724, Apr. 2021.
[4] Y. Zhang, H. Jiang, Y. Miura, C. Manning, and C. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in Proc. Machine Learning for Healthcare, 2022, pp. 2-25.
[5] P. Rajpurkar et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," Nat. Med., vol. 25, pp. 65-69, Jan. 2019.
[6] A. Hannun et al., "Cardiologist-level arrhythmia detection with convolutional neural networks," Nat. Med., vol. 25, pp. 65-69, 2019.
[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, Las Vegas, NV, Jun. 2016, pp. 770-778.
[8] X. Wang et al., "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks," in Proc. IEEE CVPR, Honolulu, HI, 2017, pp. 2097-2106.
[9] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proc. AAAI, Honolulu, HI, 2019, pp. 590-597.
[10] G. Shih et al., "Augmenting the National Institutes of Health chest radiograph dataset with expert annotations of possible pneumonia," Radiol. Artif. Intell., vol. 1, no. 1, e180041, Jan. 2019.
[11] T. Bickmore et al., "Patient and consumer safety risks when using conversational assistants for medical information," J. Am. Med. Inform. Assoc., vol. 25, no. 12, pp. 1674-1679, Dec. 2018.
[12] A. Korngiebel and S. Mooney, "Considering the possibilities and pitfalls of GPT-3 in healthcare delivery," NPJ Digit. Med., vol. 4, no. 1, p. 93, Jun. 2021.
[13] K. Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, pp. 172-180, Aug. 2023.
[14] S. Jain, A. Kulkarni, and P. Bhattacharyya, "Multilingual BERT for low-resource Indian-language clinical NER," in Proc. LOUHI Workshop, EMNLP, Abu Dhabi, 2022, pp. 55-64.
[15] W. Yim et al., "Natural language processing in oncology: A review," JAMA Oncol., vol. 2, no. 6, pp. 797-804, Jun. 2016.
[16] S. Meystre et al., "Extracting information from textual documents in the EHR," Yearb. Med. Inform., vol. 17, no. 1, pp. 128-144, 2008.
[17] Z. Obermeyer and E. Emanuel, "Predicting the future -- big data, machine learning, and clinical medicine," N. Engl. J. Med., vol. 375, pp. 1216-1219, Sep. 2016.
[18] G. Kaissis, M. Makowski, D. Rueckert, and R. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," Nat. Mach. Intell., vol. 2, pp. 305-311, May 2020.
[19] S. Tonekaboni et al., "What clinicians want: Contextualizing explainable machine learning for clinical end use," in Proc. Machine Learning for Healthcare, Ann Arbor, MI, 2019, pp. 359-380.
[20] P. Klasnja and W. Pratt, "Healthcare in the pocket: Mapping the space of mobile-phone health interventions," J. Biomed. Inform., vol. 45, no. 1, pp. 184-198, Feb. 2012.