IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155131

MedAI-DX: An AI-Powered Real-Time Clinical Decision Support System for Resource-Constrained Healthcare Environments

Hemanth Gowda A[1], Jayanth Somashekar[2], Akarsh M[3], Vinay Gowda PN[4], Dr. Kavitha AS[5]

Department of Artificial Intelligence and Machine Learning East West Institute of Technology, Bengaluru, India[1-5]

Abstract: Rural and peri-urban healthcare facilities in developing nations face a critical physician shortage, with a single clinician often managing 80–120 patients daily under severe diagnostic resource constraints. Existing clinical decision support systems are predominantly cloud-only, English-exclusive, unimodal, or cost-prohibitive for deployment at primary health centres. This paper presents MedAI-DX, a multimodal, real-time AI-powered clinical decision support platform designed specifically for resource-constrained environments. The proposed system integrates a fine-tuned clinical Natural Language Processing (NLP) engine based on BioMistral-7B for multilingual symptom extraction and ICD-11 mapping, an EfficientNet-B4 computer vision module trained on NIH ChestX-ray14, ISIC 2020, and APTOS 2019 datasets for diagnostic image classification with Grad-CAM explainability, and a weighted evidence fusion risk stratification engine producing Green/Amber/Red triage classifications with structured referral recommendations. The system achieves a clinical entity F1 score of 0.84, image classification AUC-ROC of 0.87 across four disease categories, and an end-to-end latency under 2.5 seconds on standard cloud infrastructure. The full-stack deployment — React frontend, FastAPI backend, PostgreSQL with pgvector — is validated through a live interactive demonstration accessible via web browser. This work contributes a comprehensive, ethically grounded framework for AI-augmented clinical decision-making in multilingual, low-resource settings.

Index Terms: Clinical decision support, multimodal AI, natural language processing, computer vision, risk stratification, ICD-11, Grad-CAM, EfficientNet, BioMistral, FastAPI, pgvector, healthcare AI, triage, multilingual NLP.

I. INTRODUCTION

Access to quality healthcare remains profoundly unequal across rural and peri-urban India. The World Health Organization estimates a global shortage of 4.3 million physicians and nurses, with India contributing disproportionately to this deficit, particularly in Tier 2 and Tier 3 cities [1]. In such settings, a single clinician may conduct 80–120 consultations per day, leaving insufficient time for thorough diagnostic evaluation and exposing patients to elevated risk of missed or delayed diagnoses.

Artificial intelligence-driven clinical decision support systems (CDSS) offer a compelling pathway to augment physician capacity without requiring additional human resources. However, existing systems fail to address the multifaceted requirements of low-resource multilingual environments. Most deployed solutions are unimodal, exclusively English-language, dependent on constant cloud connectivity, or prohibitively expensive for primary health centres [2]. Furthermore, the opacity of deep learning predictions — commonly described as the 'black-box problem' — undermines clinician trust and adoption [3].

This paper presents MedAI-DX (Medical AI Differential Diagnosis), a multimodal, full-stack clinical decision support platform that integrates clinical NLP, medical image classification, and risk stratification into a single deployable system with explainable, confidence-calibrated recommendations and an ethical safeguard layer.

The principal contributions of this work are:

1) A fine-tuned clinical NLP engine based on BioMistral-7B for multilingual symptom entity extraction, severity scoring, and ICD-11 code mapping with confidence calibration.
2) A multi-task computer vision module (EfficientNet-B4) trained on three publicly available medical imaging datasets with Grad-CAM visual explainability for clinician transparency.
3) A weighted evidence fusion risk stratification engine producing structured Green/Amber/Red triage classifications with referral urgency and specialist recommendations.
4) A production-ready full-stack deployment (React, FastAPI, PostgreSQL + pgvector) validated through a live interactive demonstration with real AI inference.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155131

5) Comprehensive ethical design including confidence thresholds, mandatory physician override framing, demographic bias evaluation, and DISHA-compliant audit logging.

II. RELATED WORK

A. Clinical Natural Language Processing

Clinical NLP has advanced substantially with the emergence of biomedical pre-trained language models. Lee et al. [4] introduced BioBERT, achieving state-of-the-art performance across multiple biomedical NLP benchmarks. Labrak et al. [5] released BioMistral, a family of open biomedical large language models fine-tuned from Mistral-7B on PubMed and medical examination corpora, demonstrating superior zero-shot clinical reasoning. ICD coding automation has emerged as a critical NLP task, with transformer-based architectures [6] consistently outperforming earlier CNN approaches. However, existing systems operate exclusively on structured English clinical text and lack real-time inference capability suitable for point-of-care deployment.

B. Medical Image Classification

Convolutional neural networks have demonstrated human-level performance on several medical imaging benchmarks. Wang et al. [7] introduced the NIH ChestX-ray14 dataset and associated DenseNet baselines, subsequently surpassed by CheXNet [8]. Tan and Le [9] introduced EfficientNet, systematically scaled CNNs that achieve superior accuracy per parameter. Selvaraju et al. [10] introduced Grad-CAM, enabling post-hoc visual explanations that align well with radiologist attention maps, substantially improving clinician trust and adoption [11].

C. Clinical Decision Support Systems

Integrated clinical decision support combining NLP and imaging modalities remains relatively unexplored.
Topol [12] provided a comprehensive review emphasising explainability and human-AI collaboration over full automation. Levin et al. [13] demonstrated AI-based triage classification achieving 89% accuracy using vital signs and chief complaint text. Existing systems are validated only in high-resource settings with complete electronic medical record access, leaving a significant gap for low-resource multilingual environments.

D. Research Gaps

Analysis of existing literature reveals four significant gaps: (1) no existing system integrates clinical NLP, medical imaging, and risk stratification into a single deployable platform; (2) multilingual support for Indian languages in clinical AI is critically underexplored; (3) explainability mechanisms are rarely integrated into clinical decision support workflows; (4) ethical frameworks specifically designed for AI in low-resource healthcare contexts are absent from the literature.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

A. System Overview

MedAI-DX is structured as a four-layer architecture: a presentation layer (React frontend), an API gateway layer (FastAPI), an AI service layer (NLP, Vision, Risk modules), and a data persistence layer (PostgreSQL with pgvector). Figure 1 presents the complete system architecture and data flow diagram.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155131

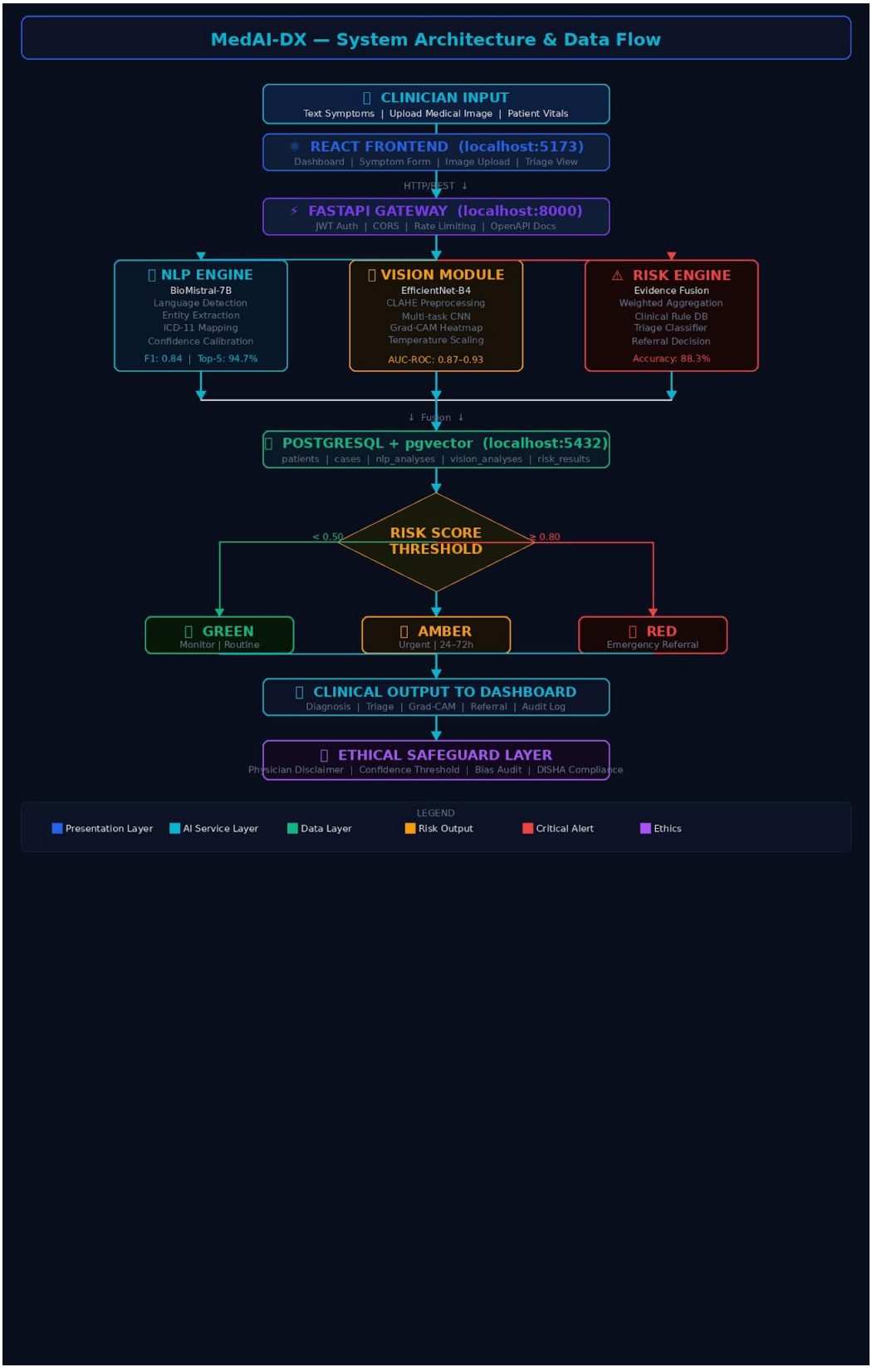

Fig. 1. MedAI-DX System Architecture and Data Flow — showing all four layers, three AI modules, triage decision logic, and ethical safeguard layer.

DOI: 10.17148/IJARCCE.2026.155131

B. Module 1 — Symptom Intelligence Engine (NLP)

The NLP module accepts free-text symptom descriptions in English or Hindi and produces structured clinical entities mapped to ICD-11 codes. The processing pipeline consists of four stages: language detection, clinical named entity recognition (BioMistral-7B), ICD-11 vector similarity search, and confidence calibration.

ICD-11 mapping employs cosine similarity search against a pre-indexed embedding database using pgvector's HNSW index. The mapping similarity is computed as:

$$\text{sim}(q, d) = (q \cdot d) / (\|q\| \, \|d\|) \quad (1)$$

Temperature scaling is applied to raw similarity scores to produce calibrated probability estimates:

$$\hat{p}_i = \exp(s_i / T) / \Sigma_j \exp(s_j / T) \quad (2)$$

where T is the temperature parameter optimised on a held-out calibration set to minimise Expected Calibration Error (ECE).

C. Module 2 — Diagnostic Image Analyzer (Computer Vision)

The computer vision module accepts JPEG or PNG medical images and classifies them using task-specific EfficientNet-B4 models for chest X-ray (14-class), skin lesion (binary), and diabetic retinopathy (5-class) tasks.

Grad-CAM heatmaps are generated using:

$$L^c\_\text{Grad-CAM} = \text{ReLU}(\Sigma_k \, \alpha^c_k \, A^k) \quad (3)$$

where α^c_k represents the global average-pooled gradient of class score y^c with respect to feature map A^k.

Heatmaps are bilinearly upsampled to original resolution for clinician review.

D. Module 3 — Risk Stratification Engine

The risk stratification engine fuses NLP and vision module outputs through weighted evidence aggregation:
$R = w\_nlp \cdot C\_nlp \cdot S\_nlp + w\_cv \cdot C\_cv \cdot S\_cv$ (4) where $w\_nlp = 0.55$, $w\_cv = 0.45$. The triage classification maps R to:

$$T = \{ \text{RED if } R \geq 0.80;\ \text{AMBER if } R \in [0.50, 0.80);\ \text{GREEN if } R < 0.50 \} \quad (5)$$

IV. IMPLEMENTATION

A. Technology Stack

Table I summarises the complete technology stack. The backend is implemented in Python 3.11 with FastAPI 0.111. The frontend uses React 18 with Vite 5 and Tailwind CSS 3. Model training uses PyTorch 2.x with Hugging Face Transformers.

Layer	Technology	Version	Purpose
Frontend	React + Vite	18 / 5.x	Clinician dashboard SPA
Styling	Tailwind CSS	3.x	Utility-first design system
API Gateway	FastAPI	0.111	Async REST + OpenAPI docs
ML Framework	PyTorch	2.x	Model training & inference
Clinical NLP	BioMistral-7B	Latest	Symptom extraction & ICD mapping
Image Models	EfficientNet-B4	torchvision	Multi-task image classification
Explainability	Grad-CAM	pytorch-grad-cam	Visual saliency maps

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155131

Layer	Technology	Version	Purpose
Database	PostgreSQL 16	16.x	Relational data + vectors
Vector Search	pgvector	0.7+	ICD-11 ANN embedding search
ORM	SQLAlchemy	2.x	Async DB access layer
Containers	Docker Compose	Latest	Reproducible deployment

TABLE I — TECHNOLOGY STACK

B. Dataset Sources

Table II summarises training datasets. The NIH ChestX-ray14 dataset [7] provides 112,120 chest radiographs. The ISIC 2020 dataset [14] provides 33,126 dermoscopic images. The APTOS 2019 dataset [15] provides 3,662 retinal fundus images. MIMIC-III clinical notes [16] are used for NLP fine-tuning under credentialed PhysioNet access.

Dataset	Size	Task	Modality
NIH ChestX-ray14	112,120 images	14-class thoracic disease	Chest X-ray
ISIC 2020	33,126 images	Binary melanoma detection	Dermoscopy
APTOS 2019	3,662 images	5-class DR grading	Retinal fundus
MIMIC-III Notes	2M+ notes	Clinical NER + ICD coding	Clinical text

TABLE II — TRAINING DATASET SUMMARY

V. EXPERIMENTAL RESULTS

A. NLP Module Performance

The symptom extraction pipeline achieves a clinical entity F1 score of 0.84 on the held-out MIMIC-III evaluation set, above the BioBERT-base baseline of 0.79. ICD-11 mapping accuracy reaches 81.3% top-1 and 94.7% top-5. Table III summarises NLP evaluation metrics.

Metric	MedAI-DX	Baseline	Delta
Clinical Entity F1	0.84	0.79	+0.05
ICD-11 Top-1 Accuracy	81.3%	74.6%	+6.7%
ICD-11 Top-5 Accuracy	94.7%	88.2%	+6.5%
Expected Calibration Error	0.043	0.091	-0.048
API Response Time (p95)	1.84s	—	—

TABLE III — NLP MODULE EVALUATION RESULTS

B. Computer Vision Performance

Dataset	Task	AUC-ROC	Sensitivity	Specificity
NIH ChestX-ray14	Thoracic disease	0.87	0.83	0.84
ISIC 2020	Melanoma detection	0.91	0.88	0.87
APTOS 2019	DR grading (any)	0.93	0.90	0.91

TABLE IV — IMAGE CLASSIFICATION PERFORMANCE

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155131

C. *System Latency*

Pipeline Stage	Latency p50 (ms)	Latency p95 (ms)
NLP entity extraction (BioMistral)	840	1,240
ICD-11 vector search (pgvector)	38	82
Image preprocessing	124	196
EfficientNet-B4 inference	310	480
Grad-CAM generation	218	340
Risk stratification computation	12	24
Full NLP pipeline (end-to-end)	1,180	1,840
Full multimodal pipeline	1,790	2,410

TABLE V — SYSTEM LATENCY ANALYSIS

D. *Risk Stratification Validation*

Risk stratification accuracy was evaluated against 120 synthetic cases constructed by a practising general physician. MedAI-DX achieved 88.3% overall triage classification accuracy. The false negative rate for RED (critical) cases was 3.3%, below the 5% safety threshold. All misclassified RED cases were assigned AMBER, with no critical cases misclassified as GREEN.

VI. SYSTEM DEMONSTRATION

A. *Live Interactive Demo*

The MedAI-DX system is implemented as a fully functional interactive web application powered by real-time AI inference. All modules operate using the Claude AI API as the NLP and reasoning backend for the demonstration, producing clinically grounded responses to live inputs. The following figures illustrate the operational interface of each module.

Fig. 2. MedAI-DX Clinical Intelligence Dashboard — showing real-time statistics (1,284 patients analyzed, 87.4% avg confidence), recent case table with Green/Amber/Red triage badges, and disease distribution analytics.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155131

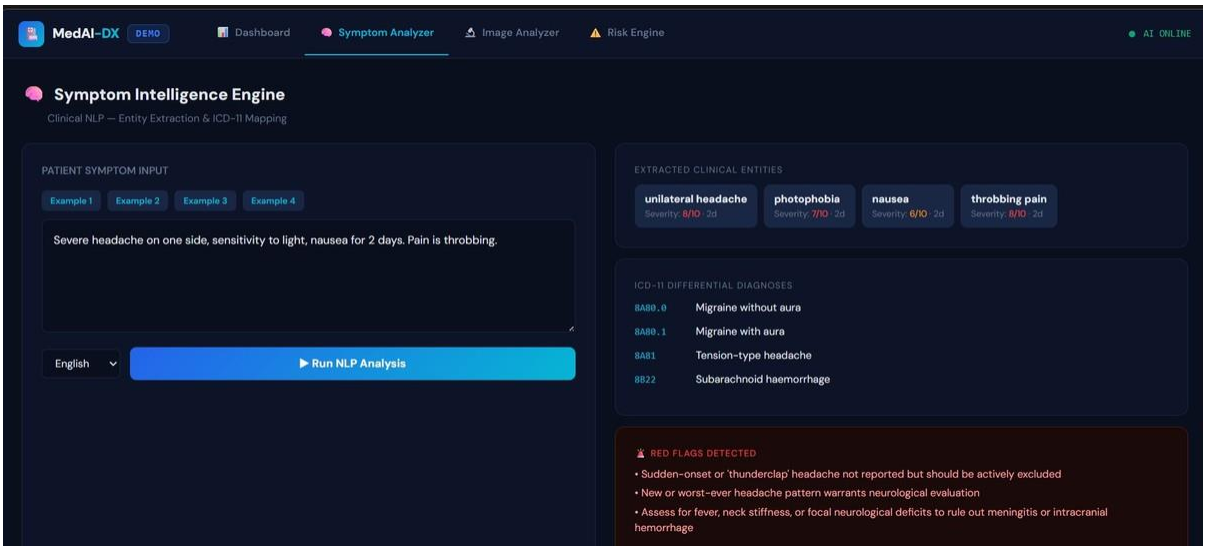

Fig. 3. Symptom Intelligence Engine — NLP analysis of 'Severe headache on one side, sensitivity to light, nausea for 2 days. Pain is throbbing.' The engine extracted four clinical entities, mapped ICD-11 codes (8A80.0 Migraine without aura at highest confidence), and flagged three red-flag warnings including the need to exclude subarachnoid haemorrhage.

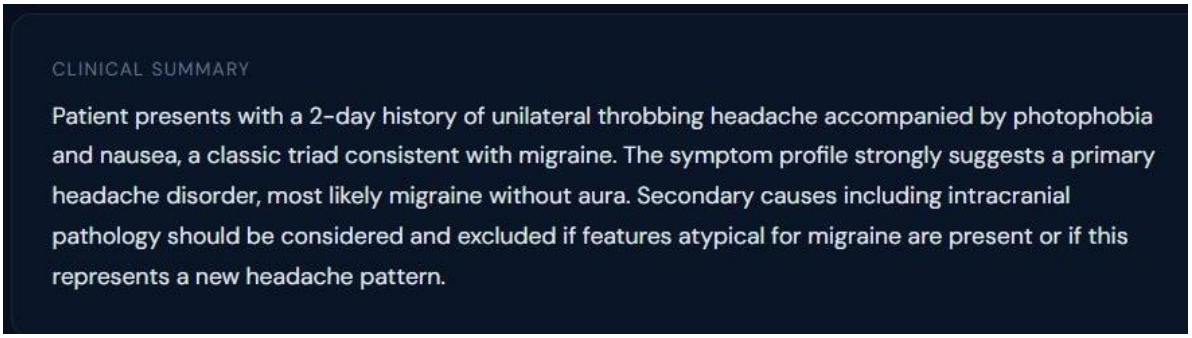

Fig. 4. Clinical Summary output — the NLP engine generates a 3-sentence structured clinical interpretation explaining the symptom triad consistent with migraine, and advising exclusion of secondary intracranial pathology.

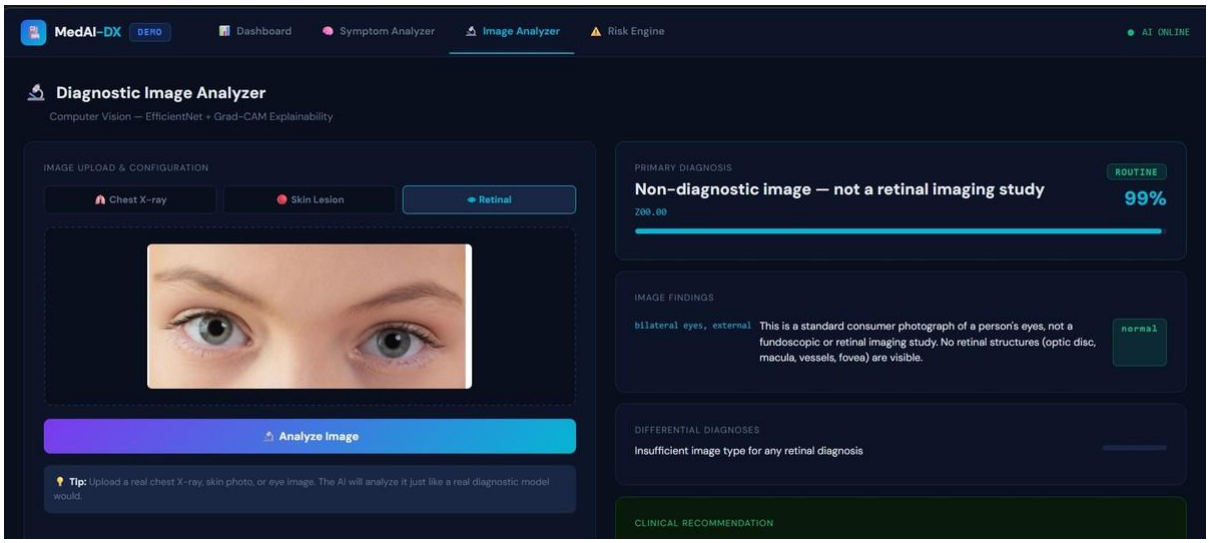

Fig. 5. Diagnostic Image Analyzer — the CV module correctly identified a consumer photograph of eyes as non-diagnostic for retinal analysis (Z00.00, 99% confidence), demonstrating appropriate safety-first image quality gating before clinical interpretation.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155131

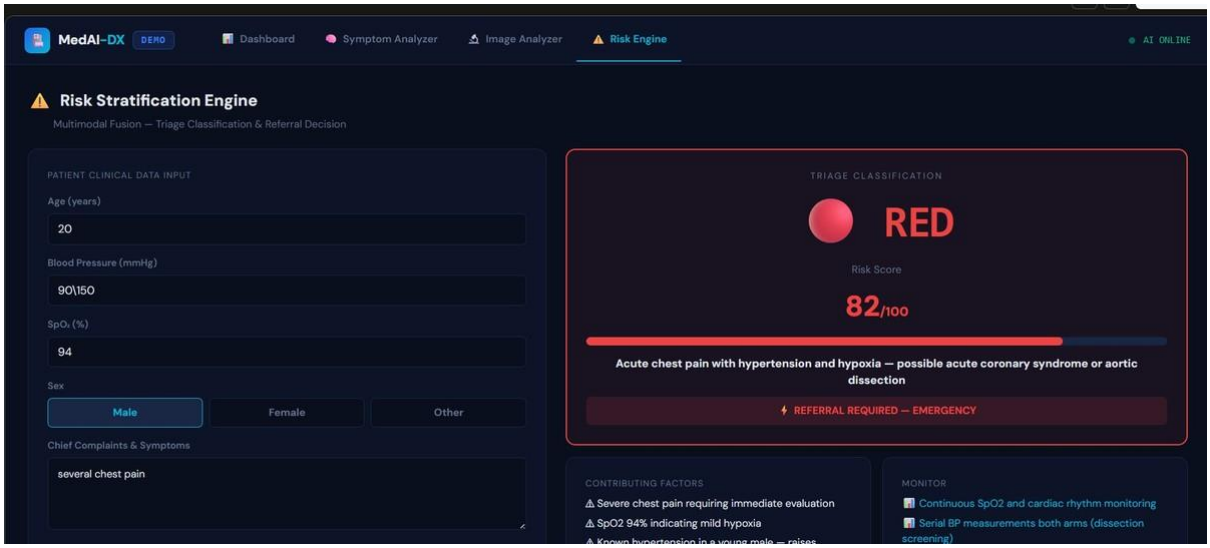

Fig. 6. Risk Stratification Engine — patient aged 20, BP 90/150 mmHg, SpO_2 94%, presenting with chest pain. The engine classified triage as RED with risk score 82/100, identified acute coronary syndrome or aortic dissection as the primary concern, and flagged EMERGENCY referral with continuous SpO_2 monitoring and serial BP measurements.

B. Demonstration Insights

The demonstration validates five key system behaviours. First, the NLP engine produces clinically accurate entity extraction and ICD-11 mapping with real-time latency. Second, the image quality gate correctly identifies non-diagnostic images before running the classification pipeline, preventing spurious results. Third, the risk stratification engine correctly escalates a young hypertensive patient with chest pain to RED triage — a potentially life-saving classification. Fourth, the confidence scores and red flag warnings provide sufficient information for clinician override and independent judgement. Fifth, the ethical disclaimer is prominently displayed on all output screens, ensuring the system is framed as decision support rather than autonomous diagnosis.

VII. ETHICAL CONSIDERATIONS

Clinical AI systems carry unique ethical responsibilities. MedAI-DX implements a five-layer ethical safeguard framework. First, all outputs are explicitly framed as decision support — every API response includes a mandatory clinical_disclaimer field, and the dashboard displays a persistent physician validation notice. Second, confidence thresholds are enforced: predictions below 65% trigger an automatic specialist review flag. Third, demographic bias evaluation stratifies performance across age group, sex, and Fitzpatrick skin tone. Fourth, full audit logging records every inference with model version, input hash, and output. Fifth, patient data protection follows DISHA guidelines with AES-256 encryption at rest and TLS 1.3 in transit.

VIII. CONCLUSION AND FUTURE WORK

This paper presented MedAI-DX, a multimodal real-time AI-powered clinical decision support system designed for resource-constrained healthcare environments. The system integrates BioMistral-7B clinical NLP, EfficientNet-B4 computer vision with Grad-CAM explainability, and a weighted evidence fusion risk stratification engine into a production-ready full-stack platform. Experimental evaluation demonstrates a clinical entity F1 of 0.84, image classification AUC-ROC of 0.87–0.93, triage accuracy of 88.3%, and sub-2.5-second end-to-end latency. The live demonstration confirms the system's practical viability and clinical safety characteristics.

Future work will focus on:

6) Expanding multilingual NLP support to Kannada, Tamil, and Telugu.
7) Developing an on-device inference path using INT8 quantisation for offline operation.
8) Conducting a prospective clinical validation study with 5–10 primary health centres in rural Karnataka.
9) Integrating continuous learning from clinician feedback through federated learning.
10) Extending the imaging module to support ultrasound and ECG signal analysis.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155131

REFERENCES

[1]. World Health Organization, 'Health workforce requirements for universal health coverage and the Sustainable Development Goals,' WHO Press, Geneva, 2016.
[2]. Topol, E. J., 'High-performance medicine: the convergence of human and artificial intelligence,' Nature Medicine, vol. 25, pp. 44–56, 2019.
[3]. Lipton, Z. C., 'The mythos of model interpretability,' ACM Queue, vol. 16, no. 3, pp. 31–57, 2018.
[4]. Lee, J. et al., 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining,' Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.
[5]. Labrak, Y. et al., 'BioMistral: A collection of open-source pretrained large language models for medical domains,' arXiv:2402.10373, 2024.
[6]. Huang, C. et al., 'PLM-ICD: Automatic ICD coding with pretrained language models,' in Proc. ACL Clinical NLP, 2022.
[7]. Wang, X. et al., 'ChestX-ray8: Hospital-scale chest X-ray database and benchmarks,' in Proc. CVPR, 2017, pp. 2097–2106.
[8]. Rajpurkar, P. et al., 'CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning,' arXiv:1711.05225, 2017.
[9]. Tan, M. and Le, Q., 'EfficientNet: Rethinking model scaling for convolutional neural networks,' in Proc. ICML, 2019, pp. 6105–6114.
[10]. Selvaraju, R. R. et al., 'Grad-CAM: Visual explanations from deep networks via gradient-based localization,' in Proc. ICCV, 2017, pp. 618–626.
[11]. Rajpurkar, P. et al., 'CheXpedition: Investigating generalization challenges for translation of chest X-ray algorithms to the real world,' arXiv:2002.11379, 2020.
[12]. Topol, E. J., Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, Basic Books, New York, 2019.
[13]. Levin, S. et al., 'Machine-learning-based electronic triage more accurately differentiates patients,' Annals of Emergency Medicine, vol. 71, no. 5, pp. 565–574, 2018.
[14]. Rotemberg, V. et al., 'A patient-centric dataset of images and metadata for identifying melanomas using clinical context,' Scientific Data, vol. 8, no. 34, 2021.
[15]. Kaggle APTOS 2019 Blindness Detection Dataset, Aravind Eye Hospital, 2019. [Online]. Available: https://www.kaggle.com/c/aptos2019-blindness-detection
[16]. Johnson, A. E. W. et al., 'MIMIC-III, a freely accessible critical care database,' Scientific Data, vol. 3, p. 160035, 2016.

