BioSpectrum: A Human-in-the-Loop Framework Combining Convolutional Neural Networks and Conversational AI for Accessible Medical Image Triage

Suyash Meena and Preenon Bagchi *

Dept. of Bioinformatics, School of Life and Basic Sciences, Jaipur National University, Jaipur, India

*Preenon.bagchi@jnujaipur.ac.in

Abstract

Access to specialist medical diagnosis remains highly unequal worldwide: the World Health Organization projects a global shortfall of roughly 11 million health workers by 2030, concentrated overwhelmingly in low- and middle-income regions [1]. Convolutional neural networks (CNNs) have repeatedly matched specialist-level accuracy on narrow, single-modality image classification tasks such as dermoscopic skin lesion classification [2], but these systems remain confined largely to research settings due to dataset, infrastructure, and validation barriers. Separately, general-purpose vision-language models (VLMs) such as GPT-4V and Claude have demonstrated an ability to reason over medical images without task-specific training, but recent evaluations show this zero-shot capability is inconsistent and, on tasks such as radiological pathology identification, has been measured at accuracies as low as 35% [3], well below what would be acceptable for autonomous diagnosis. This paper proposes BioSpectrum, a system architecture that deliberately separates these two capabilities rather than conflating them: a modality-specific, transfer-learned CNN performs the actual classification task and is evaluated against ground-truth clinical labels, while a VLM-based conversational layer explains the CNN's output, retrieves relevant recent literature, and answers follow-up questions — without altering the underlying classification. A mandatory clinician confirmation gate sits between model output and any recorded diagnosis, directly addressing the documented risk of automation bias in AI-assisted clinical decision-making [4]. We describe a two-tier deployment model: a public referral tool that routes flagged users to verified doctors and clinics without issuing a diagnosis, and a hospital-facing dashboard that supports the full clinical workflow including prescription drafting, pharmacy stock lookup, billing, and longitudinal patient records. We close with an explicit account of the system's current limitations — including dataset bias against darker skin tones documented in existing dermatology AI literature [5,6], the gap between benchmark and real-world accuracy, and the regulatory pathway required before any clinical deployment — framing these as the defined agenda for future validation work rather than as solved problems.

Keywords

Convolutional neural networks; medical image classification; vision-language models; zero-shot learning; clinical decision support; human-in-the-loop AI; automation bias; explainable AI; Grad-CAM; healthcare access; dermatology AI; telemedicine; digital health.

1. Introduction

The global distribution of medical expertise is deeply uneven. The World Health Organization estimates a worldwide shortfall of approximately 11 million health workers by 2030, concentrated in low- and lower-middle-income countries [1], and high-income countries have been reported to have more than six times as many health workers per population as low-income countries [7]. This imbalance is particularly acute in specialties that depend on visual pattern recognition —

© The Author(s) 2026

P. Bagchi et al. (eds.), *Proceedings of the International Conference on Information Systems Design And Intelligent Applications (ICISDA 2026)*, Advances in Intelligent Systems Research 215,
https://doi.org/10.2991/978-94-6239-821-4_9

dermatology, radiology, ophthalmology, and histopathology — where an estimated three billion people lack adequate access to dermatological care alone [5].

Artificial intelligence has been proposed as a partial remedy to this imbalance for over a decade. Convolutional neural networks trained on large labelled image datasets have repeatedly demonstrated diagnostic accuracy comparable to human specialists on narrow, well-defined tasks [2]. More recently, general-purpose vision-language models capable of processing both images and text — including GPT-4V, Gemini, and Claude — have shown an emergent ability to reason about medical images without any task-specific training, a property known as zero-shot inference [8].

However, treating these two classes of model as interchangeable is a mistake with real safety implications. Zero-shot VLM performance on medical images varies enormously by modality and task: one evaluation of GPT-4V on 230 emergency radiology images found correct pathology identification in only 35.2% of cases, despite correctly identifying the imaging modality itself in 100% of cases [3]. This gap between confident-sounding output and actual diagnostic accuracy is precisely the condition under which automation bias — the tendency of human operators to over-trust automated recommendations — causes the most harm [4,9].

This paper proposes BioSpectrum, a system architecture that responds to both the opportunity and the risk. Rather than using a single model for both classification and explanation, the design separates these roles: a modality-specific CNN, fine-tuned and validated on a public benchmark dataset, performs classification; a conversational VLM layer explains that output in plain language and retrieves relevant recent literature; and a mandatory human confirmation step ensures that no diagnosis is finalized without clinician review. The remainder of this paper reviews related work (Section 2), describes the proposed architecture (Section 3), discusses implementation considerations (Section 4), analyses risks and mitigations (Section 5), and states limitations and future work (Sections 6–7) prior to concluding (Section 8).

2. Related Work

2.1 Convolutional Neural Networks in Medical Image Diagnosis

Modern CNN-based image classification traces back to Krizhevsky et al.'s demonstration of deep convolutional networks on the ImageNet benchmark [10], which established transfer learning from large natural-image datasets as a practical strategy for domains with smaller labelled datasets, including medicine. Esteva et al. trained a CNN on 129,450 clinical images spanning 2,032 diseases and reported performance on par with board-certified dermatologists across both keratinocyte carcinoma and melanoma classification tasks [2], a result widely credited with establishing the feasibility of specialist-level CNN diagnosis in dermatology. Subsequent work using the International Skin Imaging Collaboration (ISIC) archive and the derived HAM10000 dataset has continued to refine this approach with architectures such as ResNet, EfficientNet, and hybrid CNN–transformer models, with reported accuracies on ISIC benchmark subsets ranging from roughly 89% to above 99% depending on the dataset subset, class count, and architecture used [11].

Despite these results, translation from benchmark performance to clinical deployment remains slow, a pattern this paper returns to in Section 2.4. The primary barriers identified in the literature are the need for large, institution-specific labelled datasets, the computational and regulatory infrastructure required for clinical-grade deployment, and the integration burden of connecting model output to existing health-record systems.

2.2 Vision-Language Models as Zero-Shot Medical Reasoners

A distinct line of work has examined whether general-purpose vision-language models — trained on broad internet-scale data rather than curated medical datasets — can perform medical image reasoning without fine-tuning. Results are mixed and strongly task-dependent. On the NEJM Image Challenge benchmark, vision-augmented large language models including GPT-4V and the Claude 3 family have been reported to rival or exceed average individual human performance, although collective human judgement from multiple clinicians still frequently outperforms any single model [12]. In radiology specifically, GPT-4V correctly identified imaging modality in essentially all cases but achieved primary pathology diagnosis accuracy of only 35.2% on a set of 230 emergency-department images [3], and a separate chest radiograph study found similarly limited performance in detecting specific radiologic findings against a radiologist-established reference standard [13]. In dermatology, zero-shot performance of ChatGPT-4o on biopsy-confirmed lesions has been shown to vary significantly by Fitzpatrick skin type, with markedly lower sensitivity and specificity for melanoma in darker skin tones [6].

Taken together, this literature supports a specific, narrower claim than the one often made in popular framing of VLMs as general medical diagnosticians: these models are useful as a conversational and explanatory layer, and are measurably unreliable as a stand-alone diagnostic classifier, particularly outside common presentations and well-represented demographic groups.

2.3 Explainability via Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM), introduced by Selvaraju et al. [14], produces a coarse localization map highlighting the image regions most responsible for a CNN's classification decision, without requiring architectural changes or retraining. The technique has since become a standard explainability tool in medical imaging AI, and empirical studies of its use have shown that Grad-CAM overlays help non-expert users distinguish a well-performing model from a poorly performing one even when both produce the same output label [14]. In dermatology AI specifically, Grad-CAM-style explainability analysis has been used to show that models trained on skin-tone-imbalanced datasets often attend to surrounding skin rather than the lesion itself when applied to darker skin tones [15], which is direct evidence that visual explanation is not merely a user-interface convenience but a diagnostic tool in its own right for surfacing dataset bias.

2.4 Automation Bias and Human-in-the-Loop Clinical AI

Automation bias — the tendency of a human operator to over-rely on automated recommendations and under-scrutinize them — is a well-documented risk in clinical decision support generally, predating the current generation of AI tools [9]. A study of 2,298 ECG interpretations found that physicians missed automated-system errors in 92 cases, and that reliance on an inaccurate system reduced internal medicine residents' own diagnostic accuracy from 57% to 48% [16]. More recent work specific to large language models has found that even AI-trained physicians remain vulnerable to automation bias when exposed to erroneous LLM recommendations, and that physicians are frequently willing to abandon an initially correct clinical decision in favour of an incorrect AI-generated one [17,18]. This literature motivates the design principle, adopted in Section 3, that a system intended for real clinical use must be structured to actively counteract automation bias rather than assume clinician vigilance by default.

2.5 Fairness and Dataset Bias in Medical Imaging AI

A separate and equally important body of work documents that medical imaging AI systems, including both trained CNNs and general-purpose VLMs, perform unevenly across patient subgroups. Daneshjou et al. found that dermatology AI models, like the dermatologists who often label their training data, show measurably reduced performance on images of dark skin tones and less common diseases, and demonstrated that fine-tuning on a more diverse dataset closed much of this performance gap [5]. Multiple subsequent analyses of the ISIC and Fitzpatrick17k datasets have quantified severe underrepresentation of Fitzpatrick skin types V–VI, with reported sensitivity drops from around 67% on a standard benchmark to as low as 11% on a dataset curated specifically for darker skin tones [15]. This body of evidence is treated in this paper not as a peripheral caveat but as a first-order design constraint, discussed further in Sections 4 and 6.

3. Proposed System Architecture

BioSpectrum is structured as a sequential pipeline with a mandatory human decision point, deployed through two distinct interfaces sharing a common backend. Figure 1 (described narratively below, as no image assets accompany this manuscript) summarizes the flow.

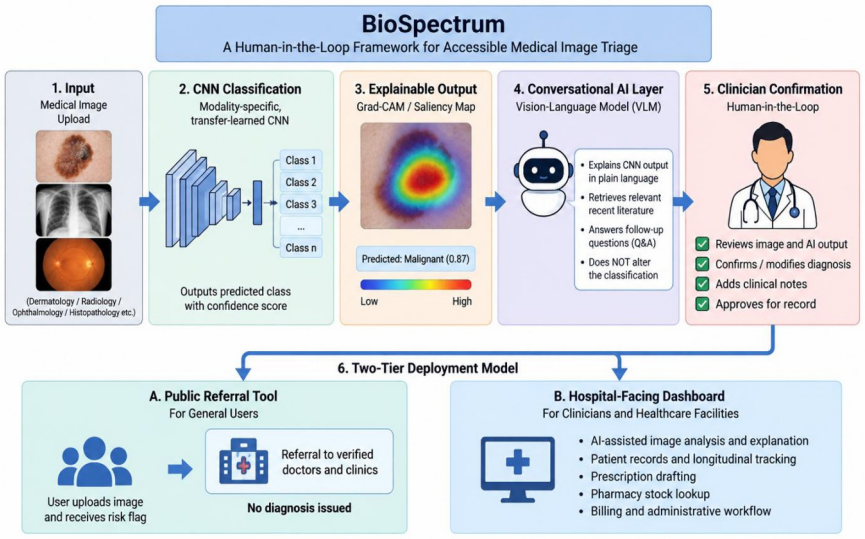


Fig. 1. Overview of the BioSpectrum system architecture. The pipeline consists of (1) medical image input, (2) modality-specific CNN classification, (3) explainable output using Grad-CAM, (4) conversational AI for explanation and literature retrieval, (5) clinician confirmation, and (6) a two-tier deployment model for public referral and hospital integration.

3.1 Stage 0: Standardized Image Capture

Image quality variance is a well-known degrader of real-world CNN performance relative to benchmark accuracy. To reduce this variance in the hospital-tier deployment, image capture is performed using a modality-appropriate attachment (for example, a dermatoscope lens or fundus camera adapter) on a fixed mount with calibrated lighting, rather than relying on unconstrained smartphone photography. An automated quality check — blur detection, exposure assessment, and region-of-interest framing — runs immediately after capture; images that fail this check are rejected and recapture is prompted, with retry counts logged per session for later audit rather than silently accepted after repeated attempts.

3.2 Stage 1: CNN-Based Diagnostic Inference

A single, clearly scoped imaging modality (for example, dermoscopic skin lesion images) is classified by a CNN built via transfer learning from an ImageNet-pretrained backbone such as ResNet or EfficientNet, fine-tuned on a public labelled dataset such as ISIC/HAM10000 [11]. The model outputs a class probability distribution and, via Grad-CAM [14], a heatmap indicating the image regions driving the classification. Validation is reported using sensitivity, specificity, and area-under-curve metrics rather than aggregate accuracy alone, given the class imbalance typical of medical imaging datasets, and performance is additionally reported per Fitzpatrick skin-type subgroup where the modality and dataset permit this, in direct response to the fairness literature discussed in Section 2.5.

3.3 Stage 2: Conversational Interpretation Layer

The CNN's structured output (class label, confidence score, Grad-CAM heatmap) is passed as context to a vision-language model, which generates a plain-language explanation for the patient or clinician, retrieves recent relevant literature via web search, and answers follow-up questions in natural language. Consistent with the evidence in Section 2.2 that VLMs are not reliable stand-alone diagnostic classifiers, this layer is architecturally prevented from overriding or re-deriving the CNN's classification; its role is strictly interpretive and communicative.

3.4 Stage 3: Mandatory Human Confirmation Gate

Before any output is recorded as a diagnosis, a clinician must review the CNN's classification alongside its Grad-CAM heatmap and either confirm or override it, entering a brief rationale in either case. This design directly targets the automation-bias risk documented in Section 2.4: rather than a single-click accept action, the interface requires active engagement with the visual evidence, and the system logs each clinician's override/acceptance rate over time so that unusually high blind-acceptance rates can be flagged for audit. The clinician's confirmed diagnosis, not the raw model output, constitutes the record of truth stored in the patient's file.

3.5 Public-Tier Referral Pathway

For public, non-clinical use, BioSpectrum deliberately restricts its output to a coarse category and urgency indicator (routine / consult soon / seek care promptly) rather than a full diagnostic report, in order to avoid the false-certainty risk documented in the VLM accuracy literature above. The urgency threshold is intentionally biased conservatively, so that borderline or low-confidence cases default toward recommending care rather than reassurance. Rather than terminating at this output, the public tier's primary function is referral: it matches the flagged condition category to relevant medical specialties and surfaces nearby doctors and clinics whose registration has been verified programmatically against a national medical registry, ranked neutrally by proximity and specialty match rather than by any paid placement.

3.6 Hospital-Tier Clinical Workflow

The hospital-facing dashboard extends the pipeline to a complete clinical episode. Following clinician confirmation of the diagnosis, a prescription may be drafted with LLM assistance but requires clinician editing and a cryptographic digital signature tied to a verified medical registration number before finalization; an automated drug-interaction check runs against the patient's recorded medication history prior to finalization. A neutral pharmacy-availability lookup shows nearby stock without paid ranking. An itemized bill is generated from the hospital's configurable fee schedule and routed through a standard payment gateway. The complete episode — image, CNN output, clinician-

confirmed diagnosis, prescription, and bill — is stored in an encrypted, access-controlled, audit-logged patient record to support longitudinal tracking across visits. A manual diagnosis-entry fallback is provided so that clinical workflow does not halt entirely if the AI pipeline or an external API becomes unavailable.

4. Implementation Considerations

4.1 Dataset, Training, and Evaluation Methodology

Consistent with common practice in medical imaging AI [2,11], the recommended implementation path fine-tunes an ImageNet-pretrained CNN backbone rather than training from randomly initialized weights, given the relatively modest size of public medical imaging datasets compared to ImageNet itself. A stratified train/validation/held-out-test split is required, with the test set untouched until final evaluation. Because benchmark datasets such as ISIC are drawn from a narrower distribution than real-world device-captured images, benchmark accuracy and prospective real-world accuracy (once device-captured data becomes available) must be reported and labelled as distinct figures rather than conflated, directly in response to the well-documented distribution-shift gap between published model performance and deployed performance.

4.2 Data Privacy and Regulatory Considerations

Any system storing identifiable patient health data in India falls under the sensitive-personal-data provisions of the Digital Personal Data Protection Act, 2023, which requires explicit consent, purpose limitation, and breach notification; where non-Indian patients or institutions are involved, HIPAA-equivalent obligations may additionally apply. Cloud storage should therefore be encrypted at rest and in transit, private by default, and configured with automated misconfiguration scanning, since misconfigured storage rather than sophisticated intrusion is the most common documented cause of health-data breaches. Software that classifies or characterizes disease from medical images is regulated as Software as a Medical Device (SaMD) in most jurisdictions, including under India's CDSCO medical device rules; a physical image-capture device of the kind described in Section 3.1 would separately fall under hardware medical-device classification. This paper treats regulatory clearance explicitly as a required future step (Section 7) rather than an implemented feature.

5. Risk Analysis and Mitigation

Table 1 summarizes the principal risk categories identified during the design of this system and the corresponding architectural mitigation, drawing on the literature reviewed in Section 2.

Risk	Mitigation
Automation bias / rubber-stamping of AI output by clinicians [4,9,17]	Mandatory active confirmation with visible Grad-CAM heatmap and logged override rates per clinician
Reduced accuracy on darker skin tones and underrepresented groups [5,6,15]	Per-subgroup performance reporting during validation; disclosure of dataset demographic composition
Zero-shot VLM misdiagnosis presented with false confidence [3,6,13]	VLM layer architecturally restricted to explanation only; CNN classification is the sole diagnostic output
Distribution shift between benchmark and real-world accuracy	Benchmark and prospective real-world accuracy reported and labelled separately; real-world validation treated as a prerequisite for clinical claims
Cloud storage misconfiguration / data breach	Encryption at rest and in transit, private-by-default storage, automated configuration scanning

Conflict of interest in pharmacy or clinic recommendations	Neutral ranking by proximity/availability only; any paid placement disclosed explicitly, never blended into ranking
Single point of failure in AI pipeline	Manual diagnosis-entry fallback preserves clinical workflow during outages

6. Limitations

This paper presents a system design and architecture, not a validated clinical tool, and several limitations should be stated plainly. First, no prospective clinical trial has been conducted; reported accuracy figures cited from the literature pertain to the underlying model classes in general, not to a specific trained instance of this system. Second, the fairness mitigations described in Section 5 reduce but do not eliminate the documented skin-tone and demographic performance gaps in dermatology AI [5,6,15]; closing this gap fully requires deliberately diversified training data, such as the DDI or Pipsqueak datasets referenced in Section 2.5. Third, the Temporal or longitudinal projections that might be built from repeated scans over time are not calibrated against longitudinal outcome data and should not be presented to patients as predictive of individual outcomes. Fourth, the system as described has not undergone regulatory review by the CDSCO, FDA, or any equivalent body, and is not authorized for autonomous clinical decision-making in any jurisdiction.

7. Future Work

- Prospective validation of the trained CNN against clinician-established ground truth on device-captured (rather than benchmark-only) images.
- Engagement with CDSCO / FDA Breakthrough Device / CE-MDR pathways appropriate to the chosen modality and risk class.
- Fine-tuning or dataset augmentation using diversified dermatology datasets (e.g., DDI, Fitzpatrick17k-balanced subsets) to close documented skin-tone performance gaps.
- HL7 FHIR-compliant integration layer to connect with existing hospital information and electronic health record systems.
- Formal usability and automation-bias testing with practicing clinicians prior to any pilot deployment.

8. Conclusion

This paper has proposed BioSpectrum, a system architecture for accessible medical image triage that deliberately separates two capabilities frequently conflated in current discourse around AI in healthcare: a validated, modality-specific CNN for classification, and a general-purpose vision-language model for explanation and communication. Grounded in the documented limitations of zero-shot medical VLM performance [3,6,13], the well-established risk of automation bias in clinical AI [4,9,17], and the persistent fairness gaps in medical imaging datasets [5,6,15], the design places a mandatory, actively-engaged clinician confirmation step between any model output and any recorded diagnosis, and separates a non-diagnostic public referral tool from a full hospital clinical workflow. The contribution of this paper is the architecture and its explicit, literature-grounded risk analysis, not a claim of validated clinical performance; Section 7 sets out the prospective validation and regulatory work required before any such claim could be responsibly made.

References

[1] World Health Organization. Health workforce. WHO projects a global shortfall of approximately 11 million health workers by 2030, concentrated in low- and lower-middle-income countries. who.int/health-topics/health-workforce

[2] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118 (2017).

[3] Waisberg, E., et al.: Assessing GPT-4 multimodal performance in radiological image analysis. European Radiology (2024).

[4] Goddard, K., Roudsari, A., Wyatt, J.C.: Automation bias – a hidden issue for clinical decision support system use. Studies in Health Technology and Informatics (2011).

[5] Daneshjou, R., et al.: Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances (2022).

[6] Housley, P.P., Samia, A.M., Motaparthi, K.: Performance evaluation of ChatGPT-4o in dermatological diagnoses across Fitzpatrick skin types. Cureus 17(7) (2025).

[7] The Lancet Global Health Editorial: Prioritising the health and care workforce shortage: protect, invest, together. Lancet Glob Health (2023).

[8] Analyzing Diagnostic Reasoning of Vision–Language Models via Zero-Shot Chain-of-Thought Prompting in Medical Visual Question Answering. Mathematics 13(14), 2322 (2025).

[9] Parasuraman, R., Manzey, D.H.: Complacency and bias in human use of automation: an attentional integration. Human Factors 52(3), 381–410 (2010).

[10] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems (2012).

[11] Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data 5, 180161 (2018).

[12] Frontiers in Medicine: Diagnostic performance of large language models on the NEJM image challenge: a comparative study with human evaluators and the impact of prompt engineering. (2025).

[13] Evaluating GPT-4V (GPT-4 with Vision) on Detection of Radiologic Findings on Chest Radiographs. Radiology (2024).

[14] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017).

[15] Ruga, T., et al.: From underrepresentation to evidence: Pipsqueak dataset and explainable analysis of skin tone bias in melanoma detection. Procedia Computer Science (2026).

[16] Censinet: Clinical Decision Support System Vendor Risk: Bias, Accuracy, and Patient Safety, citing a study of 2,298 ECG interpretations (2025).

[17] Automation Bias in Large Language Model Assisted Diagnostic Reasoning Among AI-Trained Physicians. medRxiv (2025).

[18] Physician clinical decision modification and bias assessment in a randomized controlled trial of AI assistance. Communications Medicine (2025).

[19] Balas, E.A., Boren, S.A.: Managing clinical knowledge for health care improvement. In: Yearbook of Medical Informatics, pp. 65–70. Schattauer, Stuttgart (2000).

[20] Morris, Z.S., Wooding, S., Grant, J.: The answer is 17 years, what is the question: understanding time lags in translational research. Journal of the Royal Society of Medicine 104(12), 510–520 (2011).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

