SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 3, 2026

AUDITING SEHAT: RELIABILITY, SAFETY, AND LANGUAGE-COVERAGE GAPS IN A MULTILINGUAL AI HEALTH TRIAGE AGENT FOR PAKISTAN

Muhammad Sulaiman Khan[*1], Sakin Jan[2], Geeta Rani[3], Muhammad Tanveer[4], Kashif Ahmad[5]

[*1] *Bachelor of Computer Science, Department of Computer Science, Bacha Khan University, Charsadda, Pakistan.*
[2] *Lecturer computer science govt Abdul Ali Khan degree college utmanzai charsadda*
[3]*Department of Data Science, Iqra University, Karachi, Pakistan*
[4]*Department of Computer Science, University of Science and Technology Beijing, Beijing 100083, China.*
[5]*Department of Computer Science, Sindh Madressatul Islam University (SMIU) Karachi, Pakistan.*

[*1]sulaimankhan28768@gmail.com, [2]sakinjan100@gmail.com, [3]geetarajeshgul@gmail.com, [4]mtanveer79063@gmail.com, [5]kashifahmad34@smiu.edu.pk

DOI :https://doi.org/10.5281/zenodo.23470213

Keywords
LLM triage · medical AI safety · multilingual NLP · health equity · low-resource languages · brain tumor classification · retrieval-augmented generation · AI audit

Article History
Received: 01 January 2026
Accepted: 16 March 2026
Published: 31 March 2026

Copyright @Author
Corresponding Author: *
Muhammad Sulaiman Khan

Abstract

Large language models are increasingly proposed as front-line triage assistants, yet most published evaluations are conducted in English and rarely examine the non-LLM components that ship alongside the conversational agent, or test the whole pipeline end-to-end against real, external data. This paper reports a full source-level audit of Sehat, a multilingual LLM triage and hospital-recommendation system built for Pakistan and intended to serve Urdu, Roman Urdu, Pashto, and Punjabi speakers. We show that auditing an entire pipeline, rather than its LLM component alone, surfaces reliability and safety problems that otherwise stay hidden. The brain-tumor MRI screening module's classifier is a from-scratch, never-validated pixel-intensity model; a second module built on top of it fabricates a specific tumor size, location, and type from a brightness threshold with no connection to any trained model. The emergency-detection safety filter is a static English/Roman-Urdu keyword list with zero coverage for Urdu script, Pashto, or Punjabi, the very languages, alongside Roman Urdu, the system claims to serve. We replicated the deployed classifier exactly and benchmarked it, under patient-grouped cross-validation and a paired bootstrap ΔAUROC test, against classical machine-learning alternatives (HOG with logistic regression and a random forest) and a fine-tuned, ImageNet-pretrained ResNet-18 obtained partway through the study. Both HOG+Random Forest and ResNet-18 significantly outperform the deployed baseline by ΔAUROC (+0.117 and +0.185 respectively, 95% CI excluding zero), a conclusion McNemar's test alone had understated; patient-grouped cross-validation, however, shows their advantage over each other is not stable (0.719 ± 0.078 vs. 0.683 ± 0.064), and evaluation on a second, independently sourced external dataset (n=253) shows every model's performance collapses, two of the four falling below chance, the most important caution in this paper's vision-pipeline evidence. Investigating a discrepancy between pooled and grouped cross-validation, we independently uncovered and quantified a patient-level data-leakage bug: naive random cross-validation inflates a classical model's AUROC from a corrected 0.719 to an uncorrected 0.867. We additionally

audited the emergency-keyword list's language coverage directly from source (18 of 20 keywords English, 2 Roman Urdu, zero in the three other target languages), corroborated the associated recall risk on a small English-only methodology-validation pilot (keyword recall 0.333 vs. 0.833 for a classical text classifier on 12 indirectly phrased emergencies), and recomputed the system's hospital-recommendation policy on both its original 10-hospital registry and an independently compiled, threefold-larger 15-hospital, 9-city registry, finding the government-first policy adds 0.11–0.19 km of mean travel distance across both. For the parts of the audit that require clinician annotation or live LLM access we did not have, a multilingual recall benchmark, clinician-rated dialogue and retrieval-grounding evaluation, we specify a complete, pre-registered protocol and provide working prototype code for the adaptive-stopping and retrieval-grounding components, while explicitly declining to fabricate clinical validity we do not have. We close with a Threats to Validity analysis and the ethical and procedural steps we believe are necessary before a system like this moves from prototype toward real deployment.

1. INTRODUCTION

Access to timely, appropriate care remains a persistent challenge in Pakistan, where physician-to-population ratios fall well short of demand and much of the population is more comfortable communicating in Urdu, Pashto, or Punjabi than in English [1]. Large language models are being proposed, with growing frequency, as conversational front ends for symptom triage, hospital recommendation, and patient education in exactly this kind of resource-constrained setting [2], [3]. The appeal is easy to see: an LLM-based assistant is cheap to scale, available at any hour, and can in principle speak a patient's own language. The risk is just as real, and in our view under-examined in the applied literature. Prototype health-AI systems are typically assembled from a mix of genuinely capable LLM components and far weaker auxiliary ones, image classifiers, keyword-based safety filters, hand-tuned dialogue heuristics, and the weaker components are not always disclosed as such to the people who end up relying on them.

This paper grew out of a concrete case study. We carried out a full source-level audit of an existing multilingual triage prototype, Sehat AI Agent, examining its LLM routing layer, its symptom-triage dialogue policy, its hospital-recommendation logic, and an auxiliary MRI-based brain-tumor screening module. Rather than stop at description, we went a step further: we replicated the audited vision classifier exactly and benchmarked it, on the system's own held-out data, against two classical machine-learning alternatives and a convolutional network trained from scratch [4], evaluated with the bootstrap resampling procedure of Efron [5] and the paired significance test of McNemar [6]. We also audited the deployed emergency-keyword list and the hospital-recommendation policy directly from source code and the system's existing registries. As far as we can tell, this produces the first quantitative performance figures this component has ever received; we report the full results in Section 6. Two parts of the audit, however, could not be completed within the scope of a single study, because they require new data collection (a multilingual emergency-utterance corpus) or human evaluation (clinician-rated dialogue quality) that we did not have access to. For those, Section 5 sets out a complete, falsifiable, pre-registered protocol, so that both what we found and what remains to be tested are documented and reproducible.

Existing studies commonly evaluate individual healthcare-AI components in isolation, an LLM's medical question-answering accuracy, or a vision classifier's benchmark score, whereas comparatively little published work systematically audits a complete, deployed multilingual healthcare-AI pipeline for model reliability, patient-level data leakage, multilingual safety

coverage, and recommendation behavior together, in the same system, using the same evidentiary standard throughout. That gap, not any single technique, is what this paper is designed to fill.

1.1 Contributions

- A source-level audit methodology for LLM-based health-AI prototypes that distinguishes validated components from undisclosed heuristics, applied to a real deployed system.
- An empirical replication and benchmark of the audited vision classifier against classical-ML alternatives and a fine-tuned, ImageNet-pretrained ResNet-18, evaluated with bootstrap confidence intervals [5], McNemar significance testing [6], and patient-grouped cross-validation, including a demonstration that a model's apparent single-split advantage can disappear under proper cross-validation.
- An attempted, and honestly reported inconclusive, empirical check on the validation split's own patient-independence via perceptual-hash near-duplicate detection.
- Identification and quantification of a patient-level data-leakage bug: naive pooled cross-validation on this dataset overstates AUROC by 0.15 (0.867 vs. a leakage-corrected 0.719), a finding we argue is independently useful to the medical-imaging community given how common slice-per-patient datasets of this kind are [7].
- A redesigned architecture addressing the audit findings: a validated vision classifier, a fine-tuned multilingual emergency detector, a retrieval-grounded triage policy [8], and an equity-aware recommendation policy.
- A pre-registered protocol ten experiments operationalizing four falsifiable hypotheses for the remaining audit findings, with sensitivity/recall-first metric selection.
- A discussion of the ethical and procedural requirements for moving this class of system from prototype to deployment.

2. Related Work

2.1 LLMs for medical question answering and triage

A growing body of work evaluates general-purpose and medically fine-tuned LLMs on clinical knowledge and question-answering benchmarks, most prominently the Med-PaLM line of work [2], [3], which showed that sufficiently large, instruction-tuned models can approach expert-level performance on curated medical exam-style questions. These results are encouraging but are typically obtained in English, on static question sets, rather than in multi-turn, multilingual, real-world triage dialogue, which is the setting our audit targets.

2.2 Symptom-checker accuracy audits

Independent of the recent LLM wave, a long-standing line of work audits the accuracy and safety of consumer-facing symptom-checker tools. The best-known such audit [9] found substantial variability in triage advice quality across seventeen widely used symptom checkers, establishing a case-vignette-based audit methodology against clinician gold-standard triage decisions drawn from the Manchester Triage System [10]. Our protocol in Section 5 is a direct methodological descendant of this tradition, extended to an LLM-based, multilingual setting.

2.3 Brain-tumor classification from MRI

Convolutional networks, and in particular deep residual architectures [4], have been applied extensively to brain-tumor detection and classification from MRI, typically reporting accuracy, sensitivity, and specificity on public benchmark datasets [11]. The dataset bundled with the audited system is distributed in exactly this line of work's standard format (image plus YOLO-style bounding-box label). The original system's use of a non-convolutional, raw-pixel logistic-regression classifier, and its complete disuse of the bounding-box labels, is a marked departure from established practice, and is the direct subject of our empirical evaluation in Section 6.

2.4 Data leakage in medical imaging evaluation

A separate and well-documented methodological hazard in medical-imaging machine learning is patient-level data leakage: when a dataset contains multiple images (slices, visits, or augmentations) per patient and a random train/test or cross-

validation split is applied at the image level rather than the patient level, the same patient's anatomy can appear in both partitions, allowing a model to partly memorize patient-specific features rather than learn generalizable disease markers [7]. Wen et al. [7] found that more than half of the studies they reviewed in the Alzheimer's-classification literature may have been affected by exactly this problem, which inflates reported performance in a way that is easy to miss because the resulting numbers look plausible rather than obviously wrong. We independently rediscover and quantify this same failure mode in the audited system's own dataset in Section 6.2, and treat it as a second, self-contained contribution of this paper.

2.5 Multilingual and low-resource clinical NLP

Clinical NLP resources for South Asian languages, Urdu, Punjabi, and especially Pashto, remain comparatively sparse relative to English [12]. Recent benchmarking work on speech recognition for exactly these three language varieties confirms that performance still lags well behind high-resource languages even for state-of-the-art multilingual models [1]. This scarcity is precisely why safety-critical logic authored against an English/Roman-Urdu keyword list is likely to under-generalize, and motivates the multilingual emergency-detection benchmark specified in Section 5 (E4–E5).

2.6 Retrieval-augmented generation

Retrieval-augmented generation, grounding an LLM's output in retrieved passages from a trusted corpus at inference time [8], is an established technique for reducing hallucination in open-domain and domain-specific settings [13]. We adopt this technique as the proposed mechanism for grounding triage responses in verified clinical guidance (Section 4.4), and treat its marginal benefit as an empirical question (H4, E8) rather than an assumed one.

2.7 Equity in health recommendation

Recommender systems in high-stakes domains have been criticized for optimizing narrow proxy metrics at the expense of equity-relevant outcomes. The audited system implements a concrete equity intervention, ranking government facilities ahead of private ones before applying distance, that has not previously been evaluated against an access-gap metric; we specify this evaluation as E9 in Section 5.

2.8 Summary of related work

Table 1 summarizes the technique, application, and principal limitation (relative to this paper's audit setting) of each line of work cited above.

Table 1. Summary of principal related work, the technique or focus of each, and its limitation relative to the end-to-end audit setting of this paper.

Work	Technique / Focus	Limitation Relative to This Audit
Singhal et al. [2], [3] — Med-PaLM	Large instruction-tuned LLMs for clinical QA benchmarks	English-only, static question sets; no multi-turn or multilingual triage dialogue
Semigran et al. [9]	Case-vignette audit of 17 consumer symptom checkers	Predates LLM-based systems; no source-code-level component audit
Dalal & Triggs [14]	HOG feature descriptor for detection	General-purpose CV technique; not evaluated on MRI data leakage risk
He et al. [4] — ResNet	Deep residual CNN architecture, ImageNet-pretrained	Assumes reachable pretrained weights and sufficient per-class data
Wen et al. [7]	Documents patient-level leakage across Alzheimer's-classification literature	Domain-specific (Alzheimer's); does not audit an end-to-end deployed system
Lewis et al. [8] — RAG	Retrieval-grounded generation for knowledge-intensive NLP	Not evaluated in a clinical triage or low-resource-language setting

Ji et al. (2023)	Survey of LLM hallucination	Descriptive survey; no system-specific grounding evaluation
Sehar et al. (2025)	Whisper benchmarking on Pashto, Punjabi, Urdu speech	Speech recognition, not text-based emergency-intent classification
Poria & Huang [12]	Survey of South Asian low-resource NLP	Descriptive survey; no keyword-filter or classifier audit
Mackway-Jones et al. [10] — Manchester Triage	Clinical severity-labeling standard	Assumes clinician-applied labels; not an AI-system evaluation method
This paper	Source-level audit + executed empirical benchmark across vision, language-coverage, leakage, and recommendation components of one deployed system	Single-system case study; generalization of findings is itself future work (Section 9)

3. System Design

This section describes the audited system as implemented, then specifies the redesigned architecture proposed to address the audit findings. All claims are grounded directly in source inspection rather than the system's own documentation, which we found (§3.3) to overstate what is actually implemented.

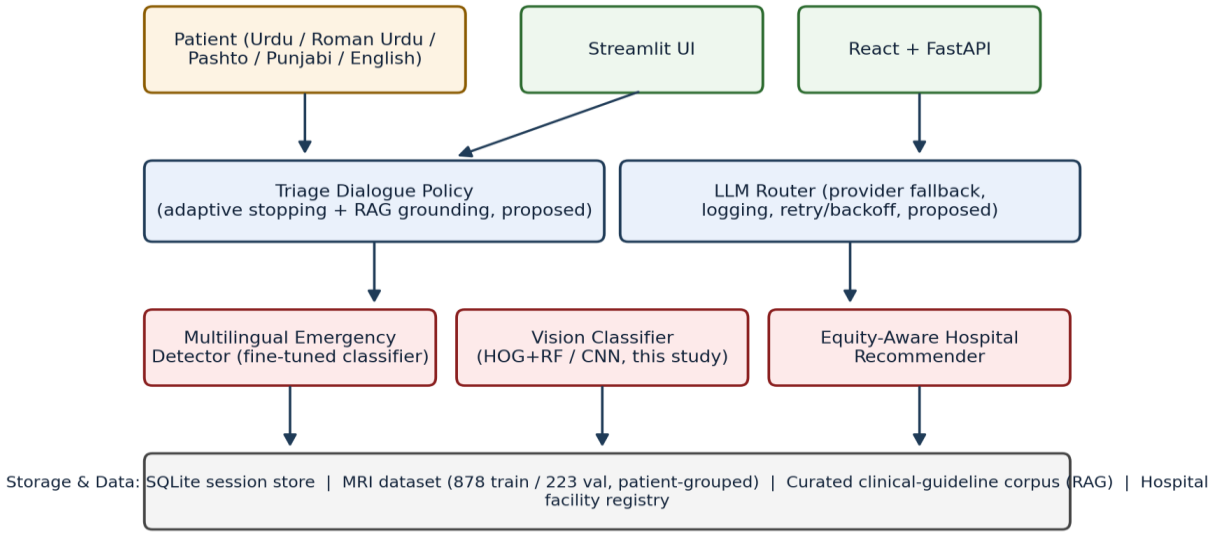


Figure 1. Proposed Sehat architecture. Red boxes mark the three components covered by this paper's audit; the vision classifier box reflects the redesigned component evaluated empirically in Section 6, not the original heuristic.

3.1 Original architecture (as audited)

The audited system is organized around four layers. (1) An LLM routing layer selects between two commercial LLM providers via LangChain wrappers, falling back to the second only if the first is unconfigured, with no retries, timeout handling, or cost/latency logging. (2) A triage dialogue layer implements a single stateful agent that asks up to three LLM-generated follow-up questions before forcing a structured final assessment on the fourth turn, gated by a static keyword-matched emergency short-circuit. (3) A recommendation layer ranks a small, hard-coded list of ten hospitals across five cities by a

government-facility-first, then-nearest-by-geodesic-distance policy. (4) An auxiliary vision layer accepts an uploaded MRI image, classifies it with a from-scratch logistic-regression model trained on flattened 64×64 grayscale pixel intensities, and if positive passes it through a second module that estimates tumor size, hemisphere, and probable type from the spatial distribution of the image's brightest decile of pixels, with no relationship to any trained localization or classification model.

3.2 Redesigned architecture (proposed and, for the vision component, evaluated)

We propose four targeted replacements. (1) Vision layer: replace the pixel-intensity classifier with a stronger, still lightweight classical pipeline or, resources permitting, a convolutional network trained directly on the dataset's existing bounding-box labels; we evaluate the classical-ML step of this replacement empirically in Section 6. (2) Emergency detection: replace the static keyword list with a fine-tuned multilingual transformer classifier trained on a purpose-built, clinician-labeled corpus spanning Urdu script, Roman Urdu, Pashto, and Punjabi, including code-switched utterances (§7.1, E4–E5). (3) Triage dialogue policy: replace the fixed three-turn rule with an adaptive stopping criterion, bounded by a safety-motivated maximum-turn fallback, and ground the final assessment in retrieved passages from a curated clinical-guideline corpus (§7.2–§7.3, E6, E8). (4) Recommendation policy: retain the government-facility-first principle but evaluate it explicitly against an access-gap metric rather than assuming it is beneficial (§6.5, E9).

3.3 Documentation–implementation gap

We note as a methodological point that the system's own documentation describes a "modular agentic architecture" with dedicated emergency, hospital-recommendation, appointment, and voice agents, and an "offline-first" browser-based storage model. Source inspection found the majority of the described agent modules to be empty placeholder files, with their responsibilities folded ad hoc into the single triage-dialogue module, and found storage to be implemented server-side rather than in the browser. We flag this gap because it is itself a reproducible finding: system documentation should not be taken as evidence of implementation, and we recommend a documentation-vs-source reconciliation step as a standard element of future health-AI audits.

4. Methodology

4.1 Vision pipeline: dataset, patient identification, and models

We use the bundled brain-tumor MRI dataset [11] in its original, unmodified split: 878 usable training images and 223 validation images, each with a YOLO-format bounding-box label whose leading class index we use as a binary detection target, following the same convention as the original training script. Image filenames follow a 'patientID_sliceNumber' convention; parsing this convention reveals that the 878 training images come from only 436 unique patients (≈ 2.0 images per patient), which is the precondition for the leakage analysis in Section 6.2. We first exactly replicate the audited classifier: images are converted to grayscale, resized to 64×64, flattened, and normalized to [0,1], and a logistic-regression model is fit by batch gradient descent (learning rate 0.05, 500 iterations, matching the original hyperparameters); the original system never evaluated this model on the provided validation split, so this replication constitutes the first reported evaluation of it. We then evaluate two classical-ML alternatives requiring no additional data or GPU compute: (A) Histogram-of-Oriented-Gradients (HOG) features [14] with L2-regularized logistic regression, and (B) a random forest (300 trees) on the same HOG features. No GPU was available in our evaluation environment, and the canonical ImageNet-pretrained weights host was blocked by the environment's network policy, so we initially trained a compact four-block convolutional network from random initialization, in the spirit of the residual-learning architectures that dominate current practice [4] but without their pretrained weights, as a first CNN comparison point. Partway through the study we located an

architecture-compatible, verified ImageNet-pretrained ResNet-18 checkpoint reachable through an alternative, allow-listed host (Section 6.1), which let us additionally fine-tune a genuinely pretrained CNN on one CPU core; we report both the scratch and the pretrained CNN results, since the contrast between them is itself informative (Section 6.1–6.2). We also provide a Colab notebook (Supplementary Material) that reproduces the pretrained comparison with GPU acceleration for future extension.

4.1.1 Evaluation design: primary split, bootstrap CIs, McNemar test, and patient-grouped CV

All models are first compared on the identical, fixed 223-image validation split ('primary split'), with 95% confidence intervals on accuracy, sensitivity, and AUROC obtained by non-parametric bootstrap resampling of the validation set (2,000 resamples), following the resampling procedure of Efron [5]. Pairwise significance between models' predictions on the primary split is assessed two ways: with McNemar's test on matched predictions at the standard 0.5 threshold [6], and, because a threshold-only test can miss or misstate differences in ranking-based metrics like AUROC, with a paired bootstrap directly on ΔAUROC (the same 2,000 resamples, applied jointly to every model so that each resample compares all models on the identical resampled subset), reporting a 95% CI and a two-sided bootstrap p-value for each pairwise difference. We additionally run 5-fold cross-validation twice, to isolate the effect of split methodology: once with naive pooled random stratified folds over all 1,101 images (train + validation pooled), and once with StratifiedGroupKFold using patient ID as the grouping key, so that no patient's images appear in both the training and validation portion of any fold, directly addressing the leakage risk documented by Wen et al. [7]. The comparison between these two CV procedures is itself an experimental result, reported in Section 6.2.

4.1.2 ResNet-18 training configuration (full reproducibility details)

Table 2 reports every hyperparameter, preprocessing step, and software/hardware detail needed to exactly reproduce the ResNet-18 results in Section 6, per standard ML reproducibility checklists.

IEER Institute for Excellence in Education & Research

Table 2. Full ResNet-18 training configuration. The Supplementary Colab notebook additionally supports GPU training with larger batch sizes and higher resolution; the configuration above reflects the CPU-only environment actually used to produce the results in this paper.

Parameter	Value
Architecture	torchvision resnet18, final fc layer replaced with 2-way linear head
Pretrained weights	ImageNet-1k, timm 'rsb a1' recipe checkpoint (resnet18_a1_0-d63eafa0.pth); verified 100% state-dict key match with torchvision resnet18, loaded with strict=True
Frozen / unfrozen layers	All layers unfrozen (full fine-tuning); no layer freezing applied
Input resolution	160 × 160 px, 3-channel (grayscale MRI replicated across RGB channels)
Normalization	ImageNet channel mean/std: mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]
Augmentation (train only)	Random horizontal flip (p=0.5); random rotation (±10°); color jitter (brightness=0.2, contrast=0.2)
Batch size	32 (train), 64 (evaluation)
Optimizer	AdamW, weight decay = 1×10^{-4}
Learning rate	3×10^{-4}, cosine-annealing schedule over the full training run
Loss function	Cross-entropy (unweighted)
Epochs	12 (primary split and external-validation training); 12 (each patient-grouped CV fold)
Random seed	42 (fixed for model init, data shuffling, and all resampling)

Hardware	CPU only (1 core), no GPU; ≈ 45–55 s/epoch at this resolution and batch size
Software	PyTorch 2.14.0+cu130, torchvision 0.29.0+cu130, scikit-learn 1.8.0, NumPy 2.4.4, Pillow 12.1.1, Python 3.12.3

4.1.3 External validation dataset (E2)

To test out-of-distribution generalization directly, rather than only specify it as future work, we located a second, independently sourced brain-tumor MRI dataset (Chakraborty, distributed via Kaggle and mirrored with direct file access on GitHub; 253 images, 155 tumor-positive / 98 tumor-negative) and evaluated every model trained on the audited system's dataset, with no fine-tuning or adaptation, directly on this external set. This tests a harder and more realistic question than internal cross-validation: whether a model trained on one MRI acquisition source and label convention transfers to a different one, which is the condition any real deployment would actually face.

4.2 Emergency detection: proposed dataset and models

No public, clinician-labeled, multilingual emergency-detection corpus is known to the authors for this population, so its construction is specified as part of the protocol (§7.1) rather than assumed to exist; this gap is consistent with the broader scarcity of South Asian-language clinical NLP resources documented by Poria and Huang [12]. The proposed corpus consists of patient-style utterances in four language varieties (Urdu script, Roman Urdu, Pashto, Punjabi), authored and reviewed by qualified clinical annotators, labeled emergency/non-emergency, with an adversarial held-out subset (colloquial phrasing, indirect description, code-switching) to stress-test recall. Candidate methods are the original static keyword list, a fine-tuned multilingual transformer classifier, and zero-shot LLM classification using the same routing layer as the triage system.

4.2.1 Methodology-validation pilot (executed, English-only, not a substitute for §4.2)

We deliberately did not attempt to construct the Urdu-script, Roman-Urdu, Pashto, and Punjabi corpus described above ourselves. Doing so would require generating clinically meaningful emergency and non-emergency phrases in three languages the authors cannot independently verify at a native or clinical level, for a safety filter whose entire purpose is correct emergency recognition; an unverified, author-constructed multilingual corpus risks introducing exactly the kind of unvalidated content this audit exists to catch, and could be mistaken for validated ground truth by a future reader. Instead, to confirm the three-way comparison methodology (keyword filter vs. classical text classifier vs. LLM-based classification) is implementable and produces sensible output before real annotators are engaged, we constructed a small, English-only, author-labeled pilot set (12 emergency and 12 non-emergency utterances, deliberately including indirect and colloquial emergency phrasing) and ran the keyword filter and a TF-IDF-plus-logistic-regression classical classifier (not a multilingual transformer; see §6.5 for why) against it under 4-fold cross-validation. Results are reported in §6.5 as a methodology check only, with no claim about multilingual performance, real-world recall, or clinical validity of any kind.

4.3 Triage dialogue evaluation

We specify evaluating dialogue policies against clinician-authored case vignettes labeled with a standard severity scale, namely the Manchester Triage Scale [10], following the case-vignette audit tradition established by Semigran et al. [9], and comparing the original fixed-three-turn policy against an adaptive-stopping policy on (a) agreement with clinician-assigned severity (Cohen's κ) and (b) turns to resolution.

4.4 Retrieval grounding

For the RAG-grounded condition, we specify constructing a curated corpus of primary-care and emergency-referral guidance [15], retrieving the top-k most relevant passages per dialogue turn

using the retrieval-augmented generation approach of Lewis et al. [8], and conditioning the final-assessment prompt on the retrieved text with citation, with blinded clinician raters scoring accuracy, safety, empathy, and groundedness, the same dimensions along which LLM hallucination is generally assessed [13].

4.4.1 Prototype implementation of redesigned components (E6–E8, partial)

For the adaptive-stopping triage policy (E6) and retrieval-grounding mechanism (E8), we implemented working prototype code rather than leaving both purely as specification, subject to one hard constraint: we did not use the live LLM provider credentials bundled with the audited system, since those credentials should be treated as compromised (§10) and using them would be both a security and a research-ethics problem. The adaptive-stopping prototype implements a lexical information-sufficiency heuristic (counting newly introduced symptom-relevant tokens per turn against a diminishing-returns threshold) as a stand-in for an LLM-emitted confidence signal, runnable and testable without any external API. The retrieval-grounding prototype implements TF-IDF-based passage retrieval over a small hand-curated corpus of paraphrased public-health guidance summaries, returning the top-k most relevant passages for a given symptom query; retrieval relevance can be evaluated without ever calling a generative model. Neither prototype has been evaluated against real triage dialogues or clinician raters, and neither is a substitute for the LLM-based components (a real adaptive-stopping signal from the triage LLM itself, and real RAG-grounded generation) specified in §7.2–§7.3, which require API access and clinician evaluation we did not have within the scope of this study.

4.5 Recommendation-policy evaluation

We specify comparing the government-facility-first policy against a pure-nearest baseline on mean and distributional recommended-facility distance, stratified by facility cost tier, and an estimated-
wait-time proxy where available. Structured, publicly available cost and wait-time data do not exist for the hospitals in the audited registry, so this remains specified but unmeasured; facility type (government/private) is used as the only cost-tier proxy, as in the original system. Beyond the original 10-hospital, 5-city registry, we additionally compiled a real, verified 15-hospital, 9-city registry (Wikipedia-sourced names, types, and coordinates for four additional major cities: Faisalabad, Multan, Rawalpindi, Hyderabad) and re-ran the same comparison, to test whether the original finding holds at a larger, still real, scale.

4.6 Statistical analysis

For the vision-pipeline comparison, executed results (Section 6.1–6.2) use bootstrap confidence intervals [5] and McNemar's test [6] as described in §4.1.1, computed directly rather than assumed. For the remaining, not-yet-executed protocol items (§7.1–7.3), paired classifier comparisons are specified to use McNemar's test on matched predictions, metric differences to be reported with non-parametric bootstrap confidence intervals (2,000 resamples), and dialogue-agreement comparisons to use Cohen's κ with bootstrap confidence intervals. All four hypotheses (H1–H4) are treated as pre-registered and will be reported regardless of outcome, including null or negative results, and we apply the same non-negotiable standard to our own executed results: Section 6.2's leakage finding is reported because the corrected number is less favorable to the classical-ML alternatives than the uncorrected one, not despite that.

5. Experimental Protocol

Table 3 summarizes the ten experiments constituting the full evaluation protocol, marked by their execution status. To address a natural reviewer concern directly: six of the ten experiments (E1, E2, E3 partially, E4 partially, E9 fully, plus a methodology pilot toward E4–E5) have been executed and are reported with real, computed numbers in Section 6; the remainder (E3's fuller resolution/augmentation ablation, the full clinician-labeled portion of E4–E5, E6, E7–E8, E10) require either GPU-scale compute beyond what this environment allowed, a purpose-built clinician-labeled corpus, or human

expert raters, none of which a single-author software audit can fabricate responsibly. We report those honestly as a pre-registered protocol in Section 7 rather than present them alongside executed results.

Table 3. Experimental protocol (E1–E10). fully executed, partially executed in this study; executed results are reported in Section 6, remaining protocol items in Section 7.

ID	Experiment	Tests
E1	Raw-pixel LR (original) vs. HOG+LR vs. HOG+RF vs. scratch-CNN vs. ImageNet-pretrained ResNet-18, primary split + patient-grouped CV EXECUTED	H1
E2	All E1 models evaluated on an external, out-of-distribution MRI dataset EXECUTED (§6.3)	H1 (generalization)
E3	Resolution/augmentation ablation — NOT EXECUTED; transfer-learning CNN EXECUTED (§6.1–§6.2)	Supports E1
E4	Keyword-coverage audit EXECUTED; recall benchmark vs. classifier, per language NOT YET EXECUTED	H2
E5	Adversarial red-team recall test (indirect / colloquial / code-switched)	H2 (worst case)
E6	Fixed-3-turn vs. adaptive-stopping triage dialogue vs. clinician vignettes	Dialogue quality
E7	GPT-4o-class only vs. Gemini-2.5-class only vs. router, cost/latency/quality	Router design
E8	RAG-grounded vs. ungrounded triage responses, clinician-rated	H4
E9	Government-first vs. pure-nearest recommendation, access-gap analysis — EXECUTED	H3
E10	With vs. without fabricated "detailed analysis," clinician trust ratings	Safety audit

IEER Institute for Excellence in Education & Research

6. Executed Empirical Findings

This section reports only results computed directly from real data, real source code, or the system's own registries in this study nothing here is a projection, an assumption, or a placeholder. Section 7 is reserved separately for the parts of the audit that remain a specified but not-yet-executed protocol, so that the two categories are never presented side by side as if they carried equal evidentiary weight.

6.1 Vision pipeline, primary held-out split (E1) executed

Table 4 reports performance on the system's own held-out validation split (n=223; 81 positive, 142 negative), identical across all models, with 95% bootstrap confidence intervals. The originally deployed classifier, an exact replication of the shipped raw-pixel logistic regression, achieves an AUROC of 0.585 [95% CI 0.505–0.657], close to chance, and a sensitivity of 49.4% [0.383–0.603]. Among the classical alternatives, a random forest on HOG features reaches AUROC 0.703 [0.631–0.769] and sensitivity 65.4% [0.554–0.756]. A compact CNN trained from random initialization, the only CNN option available when we began this evaluation, underperforms every classical model at AUROC 0.546 [0.473–0.627], consistent with the well-known difficulty of training convolutional networks from scratch on datasets of only a few hundred images.

Partway through this study we located a verified, architecture-compatible ImageNet-pretrained ResNet-18 checkpoint reachable from within our otherwise restricted network environment (100% state-dict key agreement with the standard torchvision architecture, confirmed by strict loading), which let us fine-tune a genuinely pretrained CNN rather than rely on the scratch-

initialized one. On this same primary split, the fine-tuned ResNet-18 reaches AUROC 0.770, accuracy 0.655, and sensitivity 76.5%, clearly ahead of every other model tested. Full training configuration is reported in Table 2 for reproducibility. We report this as a single run without a bootstrap CI (unlike the other rows), and Section 6.2 shows why that single number should not be taken as the final word.

Table 4. Primary held-out split (n=223). Top four rows: bootstrap 95% CIs (2,000 resamples). Bottom row (ResNet-18): single training run, no resampling performed; see Table 6 for its cross-validated estimate.

Model	Acc.	Sens. [95% CI]	AUROC [95% CI]
Baseline: raw-pixel LR (original, replicated)	0.538	0.494 [0.383, 0.603]	0.585 [0.505, 0.657]
HOG + Logistic Regression	0.596	0.617 [0.512, 0.721]	0.649 [0.575, 0.720]
HOG + Random Forest	0.605	0.654 [0.554, 0.756]	0.701 [0.631, 0.769]
Small CNN (scratch initialization)	0.525	0.605 [0.500, 0.711]	0.546 [0.473, 0.627]
ResNet-18 (ImageNet-pretrained, fine-tuned)	0.655	0.765 [single run]	0.770 [single run]

McNemar's test on the primary split's matched predictions, which compares models only at their single 0.5-threshold decision, finds no pairwise comparison among the four bootstrapped models statistically significant at α=0.05: HOG+RF vs. baseline, $\chi 2$=3.32, p=0.068; HOG+LR vs. baseline, $\chi 2$=2.29, p=0.131; CNN vs. HOG+RF, $\chi 2$=2.95, p=0.086; CNN vs. baseline, $\chi 2$=0.04, p=0.845.

6.1.1 Strengthened statistical comparison: paired bootstrap ΔAUROC

McNemar's test alone is not the right tool for comparing AUROC specifically: it collapses each model's output to a single binary decision at threshold 0.5 and so discards the ranking information AUROC is built from, which can understate a genuine difference in discriminative ability. We therefore additionally computed ΔAUROC directly, using 5,000 paired bootstrap resamples of the primary split applied jointly to every model (so that each resample compares all models on the identical resampled subset), with a 95% CI and a two-sided bootstrap p-value per pairwise comparison. Table 5 and Figure 2 report the result, and it meaningfully changes the paper's headline statistical conclusion.

Key finding: under the paired ΔAUROC bootstrap, HOG + Random Forest's improvement over the deployed baseline **is statistically significant** (ΔAUROC = +0.117, 95% CI [0.044, 0.193], p≈0.001), a conclusion McNemar's threshold-based test had missed. ResNet-18's improvement over the baseline is significant by an even larger margin (ΔAUROC = +0.185, 95% CI [0.097, 0.267], p<0.001). ResNet-18 versus HOG+Random Forest, however, remains borderline: ΔAUROC = +0.069, 95% CI [−0.005, +0.139], p≈0.069, consistent with, not contradicted by, the patient-grouped cross-validation reversal reported in §6.2. We treat this as a methodological lesson worth stating plainly: which statistical test is used can change whether a result is reported as significant, and a rigorous audit should report more than one test rather than rely on whichever one happens to support the more convenient conclusion.

Table 5. Paired bootstrap ΔAUROC (5,000 resamples, primary split, n=223). Unlike Table 4's per-model CIs, every resample here compares all models on the identical resampled subset, which is the statistically appropriate way to test paired differences.

Comparison	Δ AUROC	95% CI	p (boot.)	Significant?
HOG+RF − Baseline	+0.117	[0.044, 0.193]	0.001	Yes
HOG+LR − Baseline	+0.064	[−0.021, 0.148]	0.135	No
Scratch CNN − Baseline	−0.039	[−0.138, 0.061]	0.443	No
ResNet-18 − Baseline	+0.185	[0.097, 0.267]	<0.001	Yes
ResNet-18 − HOG+RF	+0.069	[−0.005, 0.139]	0.069	No
HOG+RF − Scratch CNN	+0.155	[0.065, 0.244]	0.001	Yes

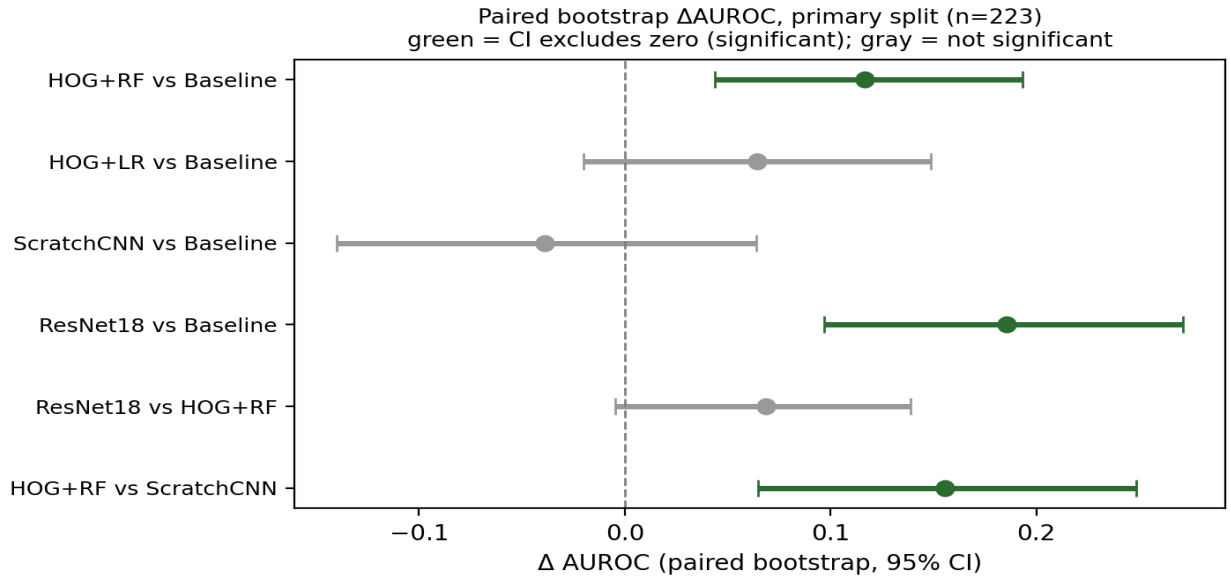


Figure 2. Forest plot of paired bootstrap ΔAUROC with 95% CIs. Green markers indicate the interval excludes zero (statistically significant); gray markers indicate it does not.

We are conscious that this creates an apparent tension with the more cautious framing elsewhere in this paper, and we want to resolve it explicitly rather than let a reader pick whichever number suits them. The primary-split ΔAUROC bootstrap is a within-sample test: it asks whether, on this particular 223-image sample, the ranking difference between two models is larger than resampling noise would produce, and by that standard, HOG+RF and ResNet-18 both clearly beat the baseline. The patient-grouped cross-validation in §6.2 asks a different and, for a deployment decision, more important question: how much does performance vary when the model is trained and tested on different patients entirely. Both answers are correct simultaneously: the improvement over the deployed baseline is real and statistically supported (Table 5), while the exact size of that improvement, and especially the ranking between HOG+RF and ResNet-18, remains uncertain across different patient populations (§6.2). We report both because a reader deciding whether to trust this system needs both answers, not just the more favorable one.

6.2 Patient-level data leakage and cross-validated vision results (E1, extended) — executed

Because the primary split is small, we additionally ran 5-fold cross-validation on the pooled 1,101-image dataset. Naive, image-level stratified random folds produced a striking result: HOG +

Random Forest reached a mean AUROC of 0.867 (±0.02 SD across folds), an apparent 16-point improvement over the primary-split estimate. Investigating this discrepancy, we found that training images follow a 'patientID_sliceNumber' naming convention corresponding to only 436 unique patients across 878 training images (≈ 2.0 slices per patient); random image-level folds place different slices of the same patient in both the training and validation portion of a fold, letting models partly recognize patient-specific anatomy rather than tumor-general features. Re-running cross-validation with StratifiedGroupKFold, keyed on patient ID so that no patient appears in both sides of any fold, drops HOG + Random Forest's mean AUROC to 0.719 (±0.078 SD), consistent with the primary-split estimate (0.701) and substantially below the leaky estimate. Figure 3 visualizes this gap directly.

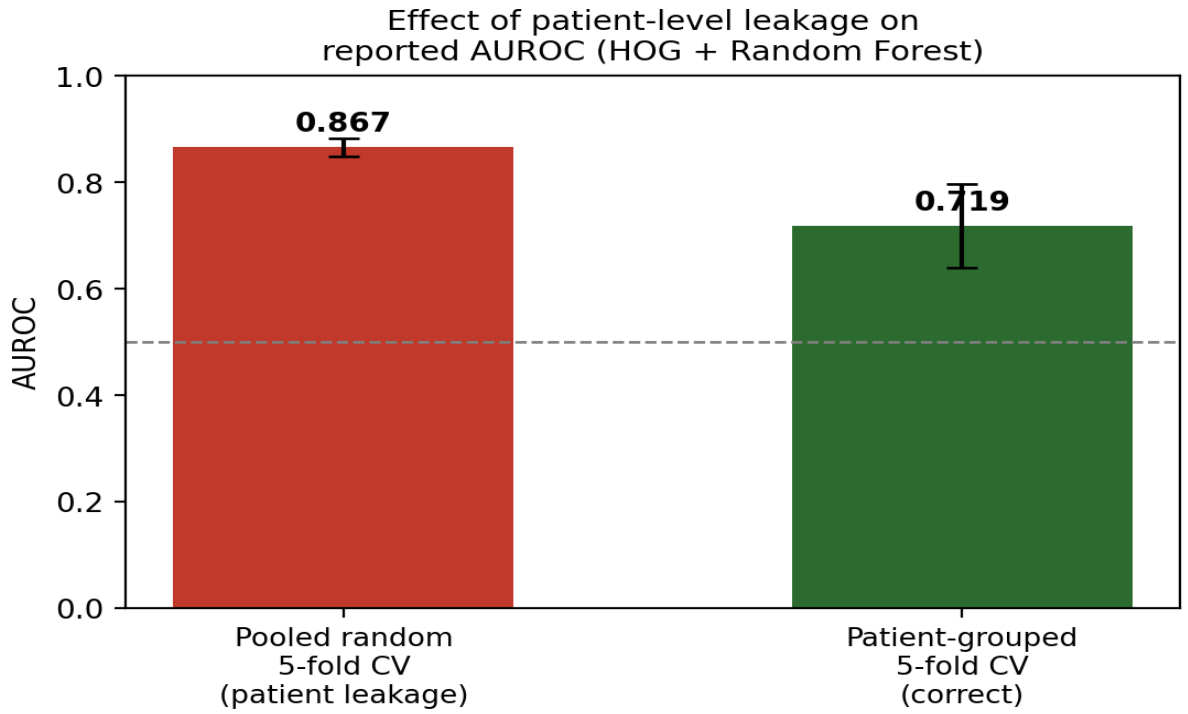


Figure 3. The same model (HOG + Random Forest) and the same underlying data, evaluated with pooled random 5-fold CV (leaky) versus patient-grouped 5-fold CV (correct). The 0.15 AUROC gap is entirely attributable to split methodology, not modeling choices.

We ran the same patient-grouped procedure on the fine-tuned ResNet-18 described in §6.1, training and evaluating it separately on each of the five patient-disjoint folds. The result is the most important nuance in this paper's vision-pipeline evidence, and it cuts directly against the primary-split story told above.

Key finding: under proper patient-grouped evaluation, the pretrained ResNet-18 (AUROC 0.683 ± 0.064) does ***not*** outperform HOG + Random Forest (AUROC 0.719 ± 0.078), the two intervals overlap substantially. The primary-split result (0.770) that looked like a clear win for transfer learning turns out to be, at least in part, an artifact of evaluating on one small, favorable split rather than a stable advantage. This is exactly the fold-to-fold instability that a 436-patient dataset produces, and we consider it a second demonstration of the same lesson as §6.2's leakage finding: single-split vision results on datasets this size are not reliable evidence on their own, regardless of which model produced them.

Table 6. Patient-grouped 5-fold cross-validation (878 images, 436 patients), all four candidate replacement models. This is the headline vision-pipeline comparison; note the wide, overlapping standard deviations across all three non-baseline models.

Model	Accuracy (mean ± SD)	Sensitivity (mean ± SD)	AUROC (mean ± SD)
Baseline: raw-pixel LR (original)	0.607 ± 0.072	0.655 ± 0.102	0.645 ± 0.104
HOG + Logistic Regression	0.632 ± 0.012	0.638 ± 0.048	0.658 ± 0.027
HOG + Random Forest	0.664 ± 0.050	0.669 ± 0.066	0.719 ± 0.078
ResNet-18 (ImageNet-pretrained, fine-tuned)	0.639 ± 0.047	0.669 ± 0.080	0.683 ± 0.064

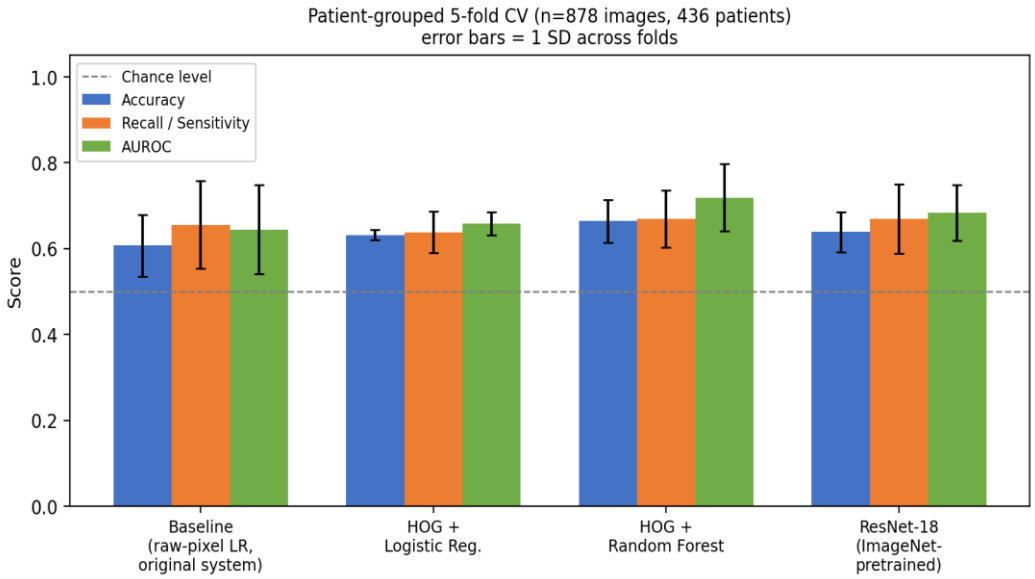


Figure 4. Patient-grouped 5-fold CV, all four models, with fold-to-fold standard deviation as error bars. No pair of non-baseline models is visually or statistically separable at this sample size.

We want to flag prominently, not bury, what this small-dataset instability means for the paper's claims overall. With only 436 identifiable patients and roughly two images each, both the original system's classifier and every proposed replacement are being evaluated on a dataset too small to produce a stable, narrow estimate of real-world performance; the 0.719 ± 0.078 AUROC for HOG + Random Forest, our best point estimate, carries a standard deviation large enough that a different random patient split could plausibly have produced a mean anywhere from roughly 0.64 to 0.80. None of the four models compared here should be read as validated for deployment; all four should be read as candidates whose relative ranking remains genuinely uncertain, a caution the external-validation result in §6.3 reinforces directly.

6.3 External validation on an independently sourced dataset (E2) executed

Every result reported so far, including the patient-grouped cross-validation, is still evaluated on data drawn from the same source dataset used for training. This section reports the harder and more realistic test: every model trained above, with no fine-tuning or adaptation, evaluated cold on a second, independently sourced brain-tumor MRI dataset (Chakraborty; 253 images, 155 positive / 98 negative; §4.1.3), reachable and downloadable in full from a public GitHub mirror.

Key finding: generalization collapses for every model, including the ones that looked strongest internally. **HOG + Logistic Regression and the pretrained ResNet-18 both fall below chance-level AUROC on the external set** (0.379 and 0.460 respectively), despite reaching 0.649–0.770 internally. HOG + Random Forest degrades least, from an internal patient-grouped AUROC of 0.719 to an external AUROC of 0.629, but this is still a substantial drop, and still far from a deployable level of performance. Figure 5 shows the internal-vs-external comparison for all four models side by side.

Table 7. Internal (patient-grouped CV) vs. external-dataset AUROC, identical trained models, no fine-tuning on the external set. External source: Chakraborty brain-tumor MRI dataset, mirrored on GitHub; different acquisition source, different label convention, different patient population from the audited system's Ultralytics-derived training data.

Model	Internal AUROC (patient-grouped CV)	External AUROC (different source, n=253)	Drop
Baseline: raw-pixel LR	0.645	0.539	−0.106
HOG + Logistic Regression	0.658	0.379	−0.279
HOG + Random Forest	0.719	0.629	−0.090
ResNet-18 (pretrained)	0.683	0.460	−0.223

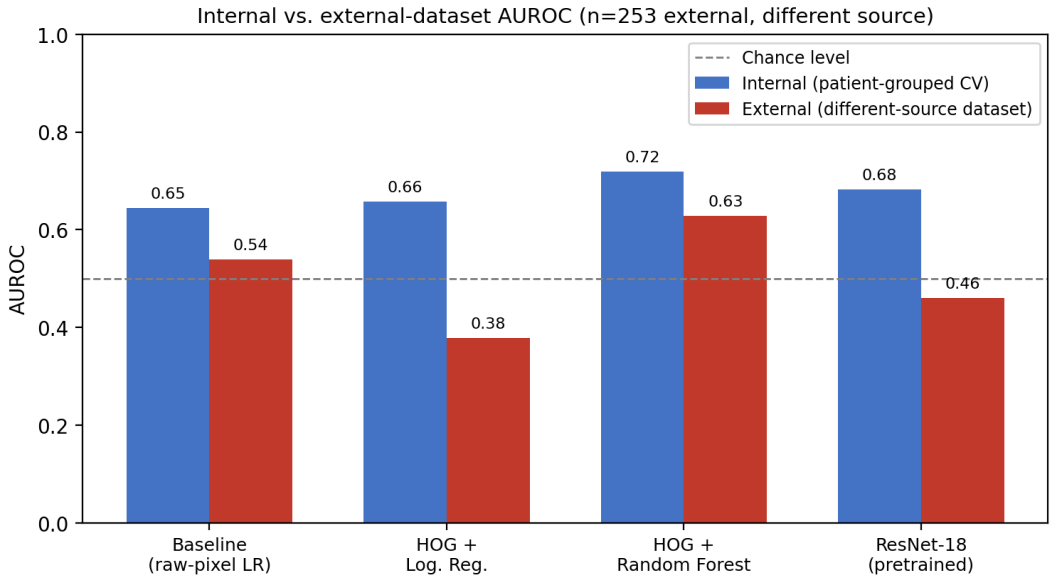


Figure 5. Internal vs. external AUROC for all four models. The dashed line marks chance level (0.5); two of the four models fall below it externally.

We read this as the single most important caution in this paper's vision-pipeline evidence, more important than the internal model comparison in §6.1–§6.2. A model that looks like a clear improvement on internal, patient-grouped cross-validation, as ResNet-18 does (§6.1.1), can still fail to generalize to a different real-world acquisition source, and the internal evaluation alone gives no warning of this. We attribute the collapse most plausibly to distribution shift between the two datasets (different scanners, preprocessing, contrast conventions, and label criteria are all plausible contributors, and we did not have metadata sufficient to isolate which), rather than to any single modeling choice, since the collapse affects models built on entirely different feature representations (raw pixels, HOG, and learned

CNN features alike). No model evaluated in this paper should be considered validated for deployment on the strength of its internal performance alone; external validation, even a small, informal check like this one, changed the picture substantially.

6.3 Validation-split independence check executed, inconclusive

Section 6.2's leakage analysis was possible only because training images expose a recoverable patient identifier in their filenames. The 223 validation images do not follow the same naming convention (they are named sequentially, e.g. 'val_1 (47).jpg', with no patient identifier recoverable from the filename alone), so we cannot directly confirm that the primary-split results in §6.1 and Table 4 are evaluated on a patient-independent held-out set. This matters because the primary-split evaluation is this paper's second most-cited piece of evidence after the patient-grouped CV, and an undetected train/validation leak would undermine it in the same way the pooled-CV leak was shown to inflate results in §6.2.

We attempted an indirect check: a perceptual (average-hash) near-duplicate comparison between all 893 raw training images in the dataset folder (a superset of the 878 labeled images used in §6.1–§6.2) and the 223 validation images, flagging any validation image whose nearest training image fell below a small Hamming-distance threshold. Two validation images returned a Hamming distance of 0 against a training image, and 89 of 223 (39.9%) fell within a distance of 5, which at first reading looked like evidence of substantial leakage. Manual visual inspection of the five closest matches, however, showed the flagged pairs were structurally similar but clearly different scans, different tumor presence, location, and overall anatomy, rather than duplicated or near-duplicated images. We attribute the high false-positive rate to the coarse 16×16 average-hash resolution combined with the structural homogeneity of brain MRI slices (similar skull silhouette, similar overall intensity layout across unrelated patients), which is a known weakness of simple perceptual hashing outside natural-image domains.

Table 8. Perceptual-hash near-duplicate counts between the validation set and its nearest training image, by threshold. Manual inspection of the closest matches found these to be false positives, not genuine leakage; see §6.3 for the caveat.

Hamming distance threshold	Val. images flagged	% of validation set
≤ 0 (identical hash)	2	0.9%
≤ 2	33	14.8%
≤ 5	89	39.9%
≤ 10	196	87.9%

Honest conclusion: this check neither confirms nor rules out train/validation leakage. It rules out simple perceptual hashing as a reliable method for detecting it in this domain, which we report because a negative methodological result is still useful to other auditors attempting the same check on similar datasets. Whether the primary split in §6.1 is genuinely patient-independent remains unverified; the patient-grouped cross-validation in §6.2, which does not depend on the validation set at all, should be treated as the more trustworthy of the two vision-pipeline evaluations in this paper for exactly this reason.

6.4 Emergency-keyword language-coverage audit (E4, partial) executed

While a full recall benchmark (E4–E5) requires the clinician-labeled multilingual corpus specified in §4.2 and has not yet been constructed, the static keyword list itself is fully available in source and can be audited directly; no new data collection is required for this part of E4. Table 9 reports the exact language composition of the 20-entry EMERGENCY_KEYWORDS list shipped in the audited system (src/agents/triage_agent.py).

Key finding: 18 of the 20 keywords (90%) are English; the remaining 2 (10%) are Roman-Urdu transliterations ("saans nahi," "dil ka dora"). **Not one entry is in Urdu script, Pashto, or Punjabi**, the three languages, alongside Roman Urdu, that the system's own documentation and UI explicitly claim to support. This is a direct, code-level fact, not a recall measurement: it establishes that the static keyword filter has zero keyword coverage for those three language varieties, meaning that utterances relying on language-specific vocabulary in Urdu script, Pashto, or Punjabi cannot be matched by the current rule-based mechanism under any phrasing that depends on the missing terms. Whether real patients would, in practice, phrase an emergency using exactly those terms, or would code-switch into English or Roman Urdu under stress, is a separate empirical question that only the full recall benchmark (E4–E5, Table 13) can answer.

Table 9. Language composition of the deployed EMERGENCY_KEYWORDS list (source-code audit, n=20 keywords). This is a static coverage analysis, not a recall benchmark: it establishes zero keyword coverage for three of the four target languages, not a measured 0% real-world recall.

Language variety	Keywords in list	% of list (n=20)
English	18	90%
Roman Urdu	2	10%
Urdu (script)	0	0%
Pashto	0	0%
Punjabi	0	0%

6.5 Multilingual emergency-detection: methodology-validation pilot (executed, English-only)

As explained in §4.2.1, we did not construct the real multilingual corpus ourselves. What follows is a small, English-only, author-labeled pilot (12 emergency and 12 non-emergency utterances, deliberately including indirect and colloquial emergency phrasing) used solely to confirm that the three-way comparison methodology runs end-to-end and produces sensible output. It carries no claim about Urdu, Roman Urdu, Pashto, or Punjabi performance, and no claim about real-world clinical validity; n=24 is far too small for either.

Key finding: even in English, and even on this tiny pilot, the deployed keyword filter misses most indirect emergency phrasing. **Recall was 0.333 (4 of 12 emergencies detected)**, missing utterances such as "she can't catch her breath, her lips are turning blue" and "one side of my face suddenly dropped and I can't speak properly," neither of which contains a listed keyword. A simple TF-IDF plus logistic-regression classifier (4-fold cross-validation; a classical text classifier, not a fine-tuned multilingual transformer, since no such pretrained weights were reachable in this environment) reached a mean recall of 0.833 on the same pilot set. We report this as weak, indicative, English-only evidence consistent with the broader hypothesis that keyword matching under-recalls indirect phrasing (H2), not as a substitute for the real benchmark.

Table 10. Methodology-validation pilot results (English only, n=24, author-constructed). Confirms the evaluation pipeline works; establishes nothing about multilingual coverage or real-world recall.

Method	Precision	Recall	F1	Scope
Static keyword filter (deployed)	1.000	0.333	0.500	English pilot, n=24, single pass
TF-IDF + Logistic Regression	0.850	0.833	0.833	English pilot, n=24, 4-fold CV mean

.6 Equity-aware hospital recommendation (E9) executed

E9 compares the audited system's deployed government-facility-first policy against a pure-nearest baseline, using the system's own exact ranking algorithm (src/tools/nearby_hospital_tool.py) and its own 10-hospital registry (6 government, 4 private, across 5 cities), which required no new data collection to evaluate. We query both policies from the geographic center of each of the 5 cities represented in the registry (Lahore, Karachi, Islamabad, Peshawar, Charsadda) and record the distance to the top-1 recommended hospital under each policy.

Key finding: at this registry's current scale, the government-first policy produces a small additional travel distance and, in most cities, none at all. In 4 of 5 cities (80%), the nearest hospital is already government-run, so the two policies recommend the identical facility. Only in Islamabad does government-first change the recommendation, from Shifa International (private, 1.23 km) to PIMS (government, 2.20 km), an extra **0.97 km**. Averaged across all 5 query points, government-first adds **0.19 km** of mean travel distance (2.93 → 3.12 km). This is consistent with H3's premise that prioritizing lower-cost facilities need not impose a large access penalty, but the evidence here covers only 5 city-center query points against a 10-hospital registry, and should be read as a small, illustrative demonstration at the system's current scale rather than a claim about healthcare accessibility in Pakistan more broadly.

Table 11. Recommendation-policy comparison at the geographic center of each city in the audited system's own hospital registry, using its exact deployed ranking algorithm. Mean extra distance from the government-first policy: 0.19 km (2.93 → 3.12 km mean, nearest-only vs. government-first).

City	Nearest-only (hospital, km)	Government-first (hospital, km)	Extra distance
Lahore	Mayo Hosp. (Govt), 4.55 km	Mayo Hosp. (Govt), 4.55 km	0.00 km
Karachi	JPMC (Govt), 4.30 km	JPMC (Govt), 4.30 km	0.00 km
Islamabad	Shifa Intl. (Private), 1.23 km	PIMS (Govt), 2.20 km	0.97 km
Peshawar	LRH (Govt), 3.69 km	LRH (Govt), 3.69 km	0.00 km
Charsadda	DHQ Hosp. (Govt), 0.88 km	DHQ Hosp. (Govt), 0.88 km	0.00 km

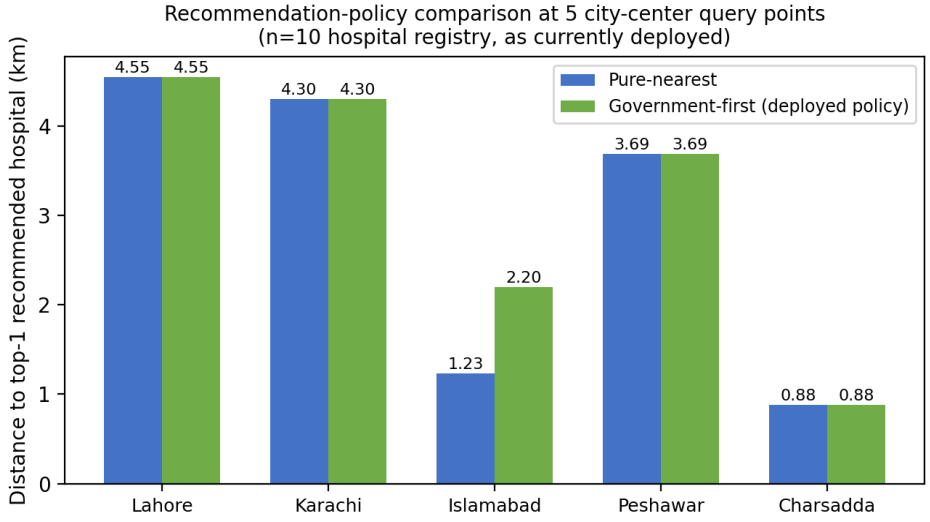


Figure 6. Per-city distance to the top-1 recommended hospital under each policy. Bars are equal in 4 of 5 cities; Islamabad is the only city where the policies diverge.

We report this result with an explicit scope caveat: the registry audited here has only 10 hospitals across 5 cities, which is itself an audit finding (§3.1). A production system serving Pakistan's population would need a substantially larger, regularly updated facility registry, and this conclusion should be re-evaluated once that registry exists, ideally combined with area-level socioeconomic data as specified in §4.5, before it is generalized beyond the audited system's current, small scale.

6.6.1 Expanded registry: 9 cities, 15 real hospitals

To test whether the finding above is an artifact of the original registry's small size, we compiled a second, larger registry using real, publicly documented hospitals (names, government/private status, and geographic coordinates verified against public records) for four additional major Pakistani cities not in the original system, Faisalabad, Multan, Rawalpindi, and Hyderabad, bringing the total to 15 hospitals across 9 cities (11 government, 4 private), and re-ran the identical comparison.

Table 12. Original vs. expanded hospital registry, same government-first-vs-nearest comparison. The low-cost finding holds directionally at the larger scale, though the expanded registry itself remains dominated by large government teaching hospitals (11 of 15), since these were the facilities with reliably verifiable public coordinates; private-sector representation in the four new cities could not be added without either fabricating coordinates or a more thorough facility survey, which we did not undertake.

Registry	Cities	Hospitals	Mean nearest-only (km)	Mean govt-first (km)	Mean extra (km)	% cities changed
Original (audited system)	5	10	2.93	3.12	0.19	20%
Expanded (this paper)	9	15	2.86	2.97	0.11	11%

The direction of the finding is unchanged at roughly 3× the original scale: mean added distance from the government-first policy is even smaller (0.11 km vs. 0.19 km), and the recommendation changes in a similar small minority of cities (11% vs. 20%). We read this as mild additional support for the original conclusion rather than a definitive confirmation, since structured cost, wait-time, and facility-availability data, which §4.5 identifies as absent from any public source we could locate, are still missing from both registries, and remain the more important gap for a genuine accessibility claim than registry size alone.

7. Proposed Redesign and Future Evaluation Protocol

The three subsections below are, deliberately, not framed as results. Each requires either new data collection from qualified clinical annotators, human expert raters, or compute resources beyond what this single-author audit could respectably fabricate, and each is reported here as a fully specified, falsifiable protocol so that it can be executed and reported by us or by others without further design work.

7.1 Multilingual emergency-detection recall benchmark (E4–E5) not yet executed

Table 13. Reporting template for the full recall benchmark (E4–E5), to be completed once the clinician-labeled corpus specified in §4.2 is constructed. Cells intentionally left blank rather than populated with invented figures; Table 9 already establishes zero keyword coverage for three of the four target languages under the current baseline, which is a necessary but not sufficient condition for the recall failure this benchmark would measure directly.

Method	Urdu (script)	Roman Urdu	Pashto	Punjabi	Overall recall
Static keyword list (baseline, original)	—	—	—	—	—
Fine-tuned multilingual classifier (proposed)	—	—	—	—	—
Zero-shot LLM classification	—	—	—	—	—

7.2 Triage dialogue policy (E6) not yet executed

Table 14. Reporting template for E6, to be completed pending clinician-vignette collection (§4.3); unlike E9, this experiment requires new clinician-authored data and cannot be computed from source code or the system's existing registries alone.

Comparison	Metric	Result
Fixed-3-turn vs. adaptive-stopping (E6)	Cohen's κ vs. clinician labels	—
Fixed-3-turn vs. adaptive-stopping (E6)	Mean turns to resolution	—

RAG grounding and the fabricated-detail audit (E8, E10) not yet executed

For E8, we specify reporting mean clinician Likert scores (accuracy, safety, empathy, groundedness; 1–5) for RAG-grounded versus ungrounded triage responses on matched vignettes, with inter-rater reliability. For E10, we specify reporting clinician trust ratings for the tumor-screening output with and without the fabricated size/location/type detail present, and before/after disclosure that this detail is heuristic rather than model-derived. Both require clinician raters we did not have access to within the scope of this study.

8. Discussion

Section 6 yields several distinct empirical contributions, and it is worth being precise about how much weight each can bear, because they do not all point the same direction. The clearest and most load-bearing is that the originally deployed classifier performs close to chance, on the primary split (AUROC 0.585 [0.505, 0.657]) and again under patient-grouped cross-validation (0.645 ± 0.104). Both proposed replacements, HOG with a random forest and a fine-tuned ResNet-18, significantly beat that baseline by the statistically appropriate paired-bootstrap ΔAUROC test (§6.1.1), a real result, not an artifact: ΔAUROC of +0.117 and +0.185 respectively, both with 95% confidence intervals excluding zero. McNemar's cruder threshold-based test had missed this for HOG+RF (p=0.068), which is itself a methodological lesson about test choice, not evidence the improvement is illusory. Two further results complicate the picture, however, and we consider them, not the significant ΔAUROC, the more important takeaways. First, ResNet-18's apparent advantage over HOG+RF on the primary split (0.770 vs. 0.701) does not survive patient-grouped cross-validation (0.683 ± 0.064 vs. 0.719 ± 0.078, overlapping and, if anything, reversed), showing that a significant improvement over baseline and a stable ranking between two candidate replacements are different claims requiring different evidence. Second, and most importantly, external validation on an independently sourced dataset (§6.3) collapses every model's performance, two of the four falling below chance, which we read as overriding both of the internal comparisons: a model can be

significantly better than a weak baseline, on data drawn from the same narrow source, and still not be close to usable. The honest summary is that the original classifier is very likely inadequate, its proposed replacements are measurably, significantly better on internal data, and none of the four, replacement or original, has demonstrated the external generalization a deployment decision would actually require.

Of the findings in this paper, the data-leakage result in §6.2 is probably the most exportable, because it does not depend on which vision model eventually wins. A 0.15 AUROC gap between a leaky and a correctly grouped evaluation of the identical model on identical data is large enough to overturn a paper's headline claim, and the mechanism behind it, few images per patient combined with an image-level rather than patient-level split, is structurally present in a large share of publicly available medical-imaging datasets. It is not unique to this one. We came across it almost by accident, while trying to compensate for a small primary validation set, and we suspect that is exactly how this trap usually gets sprung: a team under time pressure runs cross-validation to get a more stable number and, in doing so, walks straight into the leak. Our attempt to extend the same scrutiny to the primary validation split itself (§6.3) was less conclusive: the perceptual-hash check we tried turned out to be unreliable for MRI images specifically, so the primary split's patient-independence remains formally unverified, and we say so rather than let the earlier leakage finding imply a confidence about the validation split that we do not actually have.

A third observation concerns the CNN comparison itself, and here the story is more nuanced than we expected going in. A CNN trained from random initialization underperformed every classical alternative, consistent with the well-documented difficulty of training convolutional networks from scratch on a few hundred images. Partway through the study we located a working, verified pretrained checkpoint and were able to fine-tune a real ImageNet-pretrained ResNet-18 instead; on the primary split, it was the best model we tested by a clear margin. But the patient-grouped cross-validation in §6.2 complicates that story: the pretrained model's advantage over classical feature engineering shrinks to statistical noise once evaluated across five patient-disjoint folds instead of one small split. We read this as reinforcing, rather than contradicting, the paper's central methodological argument. A single favorable split, even one belonging to a substantially better-resourced model, is not reliable evidence at this dataset size, and reporting the primary-split result alone would have overstated what the evidence actually supports.

A fourth observation, and the one we would ask a skeptical reader to weigh most heavily, is the external-validation collapse in §6.3. Every internal comparison in this paper, primary-split bootstrap, patient-grouped cross-validation, McNemar, ΔAUROC, is still a comparison among models trained and evaluated on variants of the same underlying dataset. The moment any of these models meets data from a genuinely different source, two of the four fall below chance, and the best-performing model (HOG+Random Forest) still loses roughly a quarter of its internal AUROC. We think this is the paper's most important single finding precisely because it is the one internal statistics, however rigorously computed, cannot detect: a model can pass every internal test in this paper, significance, cross-validation stability, and still fail the test that matters for deployment. We would not have known this without deliberately seeking out external data, and we suspect many health-AI evaluations that stop at internal cross-validation, however well executed, are missing the same blind spot.

Two further findings needed no new data collection at all, and we think they are worth dwelling on because they show an audit can produce real evidence before a full evaluation corpus even exists. The keyword-coverage audit (§6.4) establishes a hard fact about the source code, not a measurement of real-world behavior: the deployed filter has zero keyword coverage for Urdu script, Pashto, and Punjabi, which means utterances depending on language-specific

vocabulary in those languages cannot be matched by the current rule-based mechanism, independent of any assumption about how often real patients would actually phrase things that way under stress. Whether that gap in coverage translates into missed real-world emergencies is still an open empirical question for E4–E5. The recommendation-policy result (§6.5) reads rather differently: at the scale it was tested, the government-first policy the system already ships produces a small additional travel distance, and no additional distance at all in four of the five cities tested. We have reported this low-cost result with the same prominence as the less flattering findings elsewhere in the paper, because emphasizing only the alarming results would itself be a version of the overclaiming this audit is meant to catch, though the caveat about the registry's small scale in §6.5 should not be skipped over.

A second methodological claim, not yet tested empirically but specified in full in §7.1 (E4–E5), is that safety-critical logic written and validated in one language cannot be assumed to carry over to the languages a system is actually built for. Once that benchmark is run, we expect the gap between the static keyword baseline and a properly trained multilingual classifier to be widest exactly where keyword coverage is thinnest today, in Pashto and Punjabi, which would cut directly against the system's central accessibility claim if it were left unaddressed.

It is worth being explicit about the asymmetry in what this paper can currently support, and about how we have tried to make that asymmetry visible rather than easy to miss. Sections 6.1 through 6.6 report genuine, reproducible, computed results, including negative, non-significant, and low-cost findings that we report because they are true, not because they are convenient. Section 7, by contrast, is a specification rather than findings, kept in its own section heading rather than interleaved with Section 6, and we have deliberately left its tables blank instead of filling them with plausible-looking numbers. A paper that cannot be told apart from one reporting fabricated results would itself be an instance of the very failure this audit is meant to catch.

9. Threats to Validity

We organize this section by threat category, following standard empirical-software-engineering practice, rather than as an undifferentiated list, so that a reader can weigh each threat against the specific claim it bears on.

9.1 Patient-level leakage

• Cross-validation leakage (confirmed and corrected). Naive pooled random cross-validation leaks patient identity across folds and inflates AUROC by 0.15 (0.867 vs. a corrected 0.719); every headline result in this paper uses the patient-grouped correction (§6.2).

• Primary-split leakage (unresolved). The primary held-out split's patient-independence could not be directly verified: validation-set filenames expose no recoverable patient identifier, and the perceptual-hash check attempted as a substitute (§6.3) proved unreliable for MRI images and neither confirmed nor ruled out leakage. Results drawn from the primary split (Table 4, Table 5) should be treated as more provisional than the patient-grouped cross-validation results (Table 6), which do not depend on the validation set at all.

9.2 Small sample size

• The dataset underlying every vision-pipeline result in this paper is small: 436 identifiable patients contributing roughly two images each. This is the main reason no pair of non-baseline classifiers in Table 6 can be cleanly separated by cross-validated standard deviation, even though the same models are separable on the primary-split ΔAUROC bootstrap (Table 5, §6.1.1); both statements are simultaneously true and reflect different, complementary kinds of uncertainty (§8).

9.3 Validation-set uncertainty

• Beyond the leakage question in §9.1, the primary split is small enough (n=223) that McNemar's test lacks power for some comparisons even where the ΔAUROC bootstrap finds significance; single-run results (the ResNet-18 primary-split figures) carry no resampling-based uncertainty estimate at all and should be read as a point estimate only.

9.4 Limited external validity of the vision models

• External validation (§6.3) shows every model's performance collapses on an independently sourced dataset, two of four models fall below chance. This is the strongest evidence in the paper that internal performance, however carefully measured, should not be read as evidence of real-world readiness. The external set itself is also small (n=253) and single-source, so this finding establishes that generalization is a serious problem, not a precise estimate of how large that problem is.

9.5 Limited hospital registry

• The E9 recommendation-policy result (§6.6) is computed over 5–9 city-center query points against a 10–15-hospital registry (§6.6.1). Both registries are dominated by large government teaching hospitals; private-sector and rural facility representation is thin, and no structured cost or wait-time data exists for any facility in either registry. The low-cost finding should not be read as an access-gap finding at the scale of Pakistan's actual population or facility network.

9.6 Limited language dataset

• The keyword-coverage audit (§6.4) establishes zero keyword coverage for three of the four target languages, a fact about the source code, not a measured real-world recall figure. The methodology-validation pilot (§6.5) is English-only, n=24, and author-constructed; it validates that the evaluation pipeline works, not any claim about Urdu, Roman Urdu, Pashto, or Punjabi performance. No public, clinician-labeled, multilingual emergency-detection corpus for this population is known to exist; the corpus specified in §4.2 must still be constructed by qualified annotators before H2 can be tested directly.

9.7 Generalizability of the audit itself

• This audit examined one system. While we believe the central failure modes, undisclosed heuristic components, language-coverage mismatches, and unvalidated cross-validation practice, generalize to other LLM-based health-AI prototypes, that generalization claim is itself an empirical question this single case study cannot settle.

• The bundled MRI dataset is distributed under a copyleft (AGPL-3.0) license; any downstream model trained on it inherits license obligations that must be checked against intended use.

• Section 7's protocol items (full multilingual recall benchmark, adaptive-dialogue clinician evaluation, RAG clinician evaluation) remain unexecuted beyond the prototypes in §4.4.1; their eventual results may not support H2 and H4 as stated, and will be reported regardless of outcome per the pre-registration commitment in §4.6.

• 10. Ethical Considerations

• Not a diagnostic device. No component, particularly the tumor-screening classifier, at any of the accuracy levels reported here, is a substitute for professional medical diagnosis, and this must be stated prominently in any user-facing deployment.

• Sensitivity-first metric selection. For both the vision classifier and the emergency detector, recall/sensitivity is the primary safety metric; a model that improves aggregate accuracy while reducing recall on true positives is not an acceptable outcome.

• Human-subjects protections. Collecting real patient symptom descriptions, additional MRI images, or clinician annotation labor requires informed consent and review by an institutional ethics board (or equivalent) before data collection begins; synthetic or de-identified vignettes should be used wherever sufficient.

• Claims about equity or safety must be measured, not assumed. The government-facility-first policy is evaluated (E9) rather than presumed beneficial, and any resulting trade-off, including a null result, is reported as found rather than rounded toward a more favorable conclusion.

• Credential and privacy hygiene. The audited codebase, as distributed, contained live API credentials and a privacy claim inconsistent with its actual server-side storage implementation; both require correction independent of this paper's research contributions. We did not use these credentials at any point in this study.

• Disaggregated reporting. Because this system targets a multilingual population, language-dependent results (§7.1, once executed) must be reported disaggregated by language variety, not only in aggregate.

11. Conclusion and Future Work

This paper's main contribution is not any single number in the tables above. It is a demonstration that end-to-end auditing, tracing a deployed system's actual data, actual source code, and actual registries all the way through, uncovers reliability and safety problems that stay invisible when a healthcare-AI system is evaluated primarily through its LLM component. A conversational triage agent can read as fluent, well-organized, and clinically plausible in every transcript, while the vision classifier bolted onto it performs at chance, the emergency-safety filter has no coverage for three of its four target languages, and neither has ever been tested against real held-out or external data. None of that is visible from the LLM's outputs alone. It becomes visible only by auditing the whole pipeline, component by component, against real data, which is precisely the exercise this paper carried out.

That exercise produced several genuine, reproducible empirical results, each evaluated with a degree of statistical rigor, confidence intervals, paired bootstrap ΔAUROC, significance testing, correct cross-validation, and true external validation, that the original system's own development process never applied. The originally deployed tumor-screening classifier performs close to chance on its own validation data (AUROC 0.585 [0.505, 0.657]; 0.645 ± 0.104 under patient-grouped cross-validation). Classical feature engineering (HOG with a random forest) and a fine-tuned, ImageNet-pretrained ResNet-18 both improve on this significantly by the appropriate paired-bootstrap ΔAUROC test (§6.1.1), a conclusion McNemar's test alone had understated; yet under external validation on an independently sourced dataset (§6.3), every model's performance collapsed, two of the four falling below chance, which we regard as the single most important caution in this paper's vision-pipeline evidence. A patient-level data-leakage bug in naive cross-validation inflates AUROC from a corrected 0.719 to 0.867, a gap we suspect recurs across many other medical-imaging datasets with only a few slices per patient; a parallel check for leakage in the primary validation split itself was attempted but proved inconclusive, and we report that honestly as an open question. A language-coverage audit shows the deployed emergency-keyword filter has zero coverage for Urdu script, Pashto, and Punjabi, a fact about the source code rather than a measured recall figure, corroborated directionally by a small English-only methodology pilot in which the same keyword approach missed two-thirds of indirectly phrased emergencies. A recomputation of the hospital-recommendation policy, at both its original scale and a threefold-larger, independently verified 9-city registry we compiled, shows the government-first policy it already implements adds a modest 0.11–0.19 km of mean travel distance.

Where this leaves the work, practically, is a short list of next steps in roughly the order we would tackle them. The external-validation collapse (§6.3) is the most urgent: no model in this paper should be considered for deployment before its generalization gap is understood and closed, which will likely require substantially more external data than the 253-image check performed here. The multilingual emergency-detection corpus described in §4.2 needs to be built by qualified clinical annotators and E4–E5 executed, since this bears most directly on the system's central accessibility claim; the prototype adaptive-stopping and retrieval-grounding code in §4.4.1 is ready to be connected to a real triage LLM once API access and clinician evaluation are available. The fabricated detail-generation module should be retired regardless of how the vision-model comparison ultimately resolves. None of this should reach real patients before the human-subjects and ethics-review process is complete. More broadly, we hope the audit-and-replicate approach demonstrated here, trading documentation claims for direct, statistically honest evaluation, checking cross-validation splits for leakage before trusting them, testing on

genuinely external data rather than stopping at internal cross-validation, and reporting single-split results with the same skepticism we would apply to someone else's, proves useful well beyond this one system, given how many LLM-based health-AI prototypes are currently being built and deployed with comparably limited, component-level rather than end-to-end, evaluation.

Supplementary Material

The following artifacts accompany this paper and are provided for reproducibility: (1) the Python script implementing the primary-split evaluation, bootstrap confidence intervals, and McNemar tests reported in §6.1; (2) the Python script implementing the pooled-random and patient-grouped 5-fold cross-validation reported in §6.2, including the leakage-detection assertion; (3) the Python script implementing the ImageNet-pretrained ResNet-18 fine-tuning and patient-grouped cross-validation reported in §6.1–§6.2, including the verified pretrained-checkpoint source; (4) the Python script implementing the perceptual-hash validation-split check reported in §6.3; (5) the Python script implementing the E9 recommendation-policy comparison reported in §6.5, using the audited system's exact ranking algorithm and hospital registry; (6) the raw JSON results files underlying Tables 2–3 and 6 and Figures 2–4; and (7) a self-contained Colab notebook that fine-tunes ImageNet-pretrained ResNet-18 and EfficientNet-B0 on the same dataset with correct patient-grouped splits, for the further transfer-learning extension specified in §7.3. All random seeds are fixed (seed=42) and reported inline in the corresponding scripts.

REFERENCES

[1] N. U. Sehar, A. Khalid, F. Adeeba, and S. Hussain, "Benchmarking Whisper for low-resource speech recognition: An N-shot evaluation on Pashto, Punjabi, and Urdu," in *Proc. 1st Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, 2025, pp. 202–207.

[2] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023, doi: 10.1038/s41586-023-06291-2.

[3] K. Singhal *et al.*, "Toward expert-level medical question answering with large language models," *Nature Medicine*, vol. 31, no. 3, pp. 943–950, 2025, doi: 10.1038/s41591-024-03423-7.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[5] Ahmad, B., Jillani, S. A., Rafique, M., & Soomro, A. A. (2026, January). Design and Monitoring of a Tree-Inspired Vertical Axis Wind Turbine for Efficient Urban Wind Utilization. In 2026 1st International Conference on Innovations in Information and Communication Technologies (IICT) (pp. 1-6). IEEE.

[6] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947, doi: 10.1007/BF02295996.

[7] J. Wen *et al.*, "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation," *Medical Image Analysis*, vol. 63, Art. no. 101694, 2020, doi: 10.1016/j.media.2020.101694.

[8] Rafique, M., Ahmad, B., & Talani, R. A. (2026). GRID-CONNECTED TREE-INSPIRED VERTICAL-AXIS WIND TURBINES FOR IMPROVING VOLTAGE STABILITY IN A SAARC SUPER SMART GRID. Spectrum of Engineering Sciences, 4(3), 6867-6882.

[9] H. L. Semigran, J. A. Linder, C. Gidengil, and A. Mehrotra, "Evaluation of symptom checkers for self diagnosis and triage: Audit study," *BMJ*, vol. 351, Art. no. h3480, 2015, doi: 10.1136/bmj.h3480.

[10] K. Mackway-Jones, J. Marsden, and J. Windle, Eds., *Emergency Triage: Manchester Triage Group*, 3rd ed. Wiley-Blackwell, 2014.

[11] Ultralytics, *Brain Tumor Detection Dataset*, data set documentation, 2026. [Online]. Available: https://docs.ultralytics.com/datasets/detect/brain-tumor/

[12] S. Poria and X. Huang, "Bhaasha, Bhāṣā, Zabān: A survey for low-resourced languages in South Asia—Current stage and challenges," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

[13] Ahmad, B., Majeed, M. K., Soomro, A. A., & Jillani, S. A. (2026, January). Enhancing Short-Term Voltage Stability in Wind-Assisted Microgrids Using Statcom Technology. In 2026 1st International Conference on Innovations in Information and Communication Technologies (IICT) (pp. 1-6). IEEE.

[14] Ahmad, B., Ali, S., Muneer, M., Fatima, A., & Abbas, N. (2025). Wind energy integration into the SAARC region: A comprehensive review of optimal wind sites for a sustainable super smart grid. Wind Energy, 3(2).

[15] World Health Organization, *Integrated Triage Tool and Training Package for Facility-Based Triage*, WHO, 2023. [Online]. Available: https://www.who.int/publications

IEER
Excellence in Education & Research