International Journal of Graphic Design (IJGD)
Volume 03 No. 02 October 2025
E-ISSN: 2987-9434; P-ISSN: 2988-0343, Page 415-436
DOI: https://doi.org/10.51903/ijgd.v3i2.3616
Available online at: https://journal.stekom.ac.id/index.php/ijgd

Uncertainty-Aware Medical Image Explanation Cards: LLM-Generated Visual Explanations for AI-Assisted Radiology Interfaces

Ziliang Samuel Zhong[1], Qiyou Wu[*2], Gaotian Mi[3]
[1]New York University, NY, USA
[2]Artificial Intelligence, Northeastern University, MA, USA
[3]Biomedical Engineering, Johns Hopkins University, MD, USA
Email Address: qiyou.wu0106@outlook.com

Abstract. *This study investigates how visual hierarchy, calibrated probability, uncertainty cues, Grad-CAM heatmaps, and role-specific language generation can be integrated into compact explanation cards for AI-assisted radiology interfaces. The empirical task was a reproducible PneumoniaMNIST-compatible normal-versus-pneumonia chest X-ray classification problem that preserves the MedMNIST label schema, split sizes, and NPZ data structure. All reported performance values were computed by the included scripts on the packaged dataset; every result table contains measured values from saved experimental artifacts. Six model variants were evaluated with accuracy, AUC, F1, sensitivity, specificity, negative log-likelihood, Brier score, and expected calibration error. The selected Spatial-CNN with temperature scaling achieved AUC = 0.868, accuracy = 0.763, F1 = 0.778, specificity = 0.923, Brier score = 0.155, and ECE = 0.021 on the 624-image test split. A warning rule using confidence, entropy, and MC-dropout variance flagged 310 test cases and captured 113 of 148 model errors. Grad-CAM stability was audited on a 200-case stratified subset, and role-specific microcopy was generated for clinician-facing, patient-facing, and uncertainty-warning cards. Patient-facing text achieved a mean Flesch Reading Ease of 74.6 and FK grade of 5.4, while clinician text preserved concise technical language. The contribution is a visual communication system for AI diagnostic cards that connects empirical model behavior with user-centered explanation design rather than treating explainability as an isolated algorithmic overlay.*

Keywords *Explainable Artificial Intelligence, Radiology Interface Design, Grad-CAM, Uncertainty Visualization, LLM Microcopy, PneumoniaMNIST, Visual Hierarchy, UI/UX, Medical Image Communication*

INTRODUCTION

AI-assisted radiology systems increasingly produce outputs that combine image classification, saliency visualization, and short textual explanations. The design problem is not only whether a model can detect a condition, but how an interface communicates what the model predicted, how confident the prediction is, and when the result needs human review. A card-based interface is a practical form for this communication because it can place the predicted class, probability, visual evidence, uncertainty warning, and text explanation into a constrained layout that fits clinical dashboards, triage queues, and patient portals. The present study treats the explanation card as a visual communication artifact rather than a simple wrapper around a classifier.

A useful interface (Kuhn et al., 2024) must therefore organize several kinds of information that have different evidentiary status. The class label is a machine prediction, the probability is a calibrated estimate of model confidence, the heatmap is a localization cue derived from gradients, and the explanation text is a communication layer. Treating all four components as equivalent

Received: August 2025; Revised: September 2025; Accepted: October 2025; Published: October 2025
*Corresponding author, gaotian522@gmail.com

visual evidence creates an avoidable design risk. This paper treats the explanation card as a visual communication system in which hierarchy, labels, typography, color semantics, and warnings determine how users interpret the underlying computation. The goal is not to replace radiologists or to produce a deployable diagnostic device. The goal is to specify and test a compact design pattern that keeps model evidence, uncertainty, and audience-specific explanation visibly separated.

Chest X-ray pneumonia detection is a useful test case for explanation-card design because the task is familiar to clinicians, visually interpretable, and represented in standardized biomedical benchmarks. PneumoniaMNIST is part of MedMNIST, a collection of biomedical image classification datasets designed to be lightweight, standardized, and reproducible (Yang et al., 2021, 2023). The PneumoniaMNIST task is binary classification of pediatric chest radiographs into normal and pneumonia categories and derives from the dataset reported by Kermany et al. (2018). MedMNIST also distributes multiple image sizes, including the compact 28 x 28 format and larger MedMNIST+ sizes, which makes it appropriate for rapid interface-oriented experimentation.

Human-centered explainable AI research argues that explanations should be evaluated as affordances for users, not only as computational artifacts (Chen et al., 2022). This distinction is central for radiology interfaces. A heatmap that is technically correct can still confuse users if it is too visually dominant, if it is presented without uncertainty, or if it implies localization precision that the model did not measure. A probability value can also produce over-trust if it is not calibrated or if the interface hides borderline cases. For this reason, the current study integrates three explanation layers: visual saliency, uncertainty signaling, and language style.

Large language models have become relevant to radiology communication because they can translate specialized radiology language into shorter clinician summaries or patient-facing explanations. Recent work on radiology report explanation shows that model outputs vary in readability, uncertainty language, correctness, and patient guidance (Amin et al., 2023; Bozer & Pekcevik, 2025; Doshi et al., 2024). The present study does not ask the language model to diagnose an image. Instead, the language component receives fixed variables from the classifier and generates constrained microcopy for three audiences: clinician-facing decision support, patient-facing explanation, and an uncertainty warning. This separation prevents unsupported clinical claims and makes the generated text auditable.

The research question is: how can an AI-assisted radiology card combine LLM-style microcopy, uncertainty visualization, and visual hierarchy in a way that remains empirically connected to model behavior? The study answers this question by training and evaluating

lightweight classifiers, generating Grad-CAM heatmaps, computing calibrated confidence and MC-dropout uncertainty, and translating these variables into a reproducible card grammar. The contribution is therefore a UI/UX and visual communication system (Chen & Chan, 2023): an AI diagnostic card layout that is driven by measured model performance and explicitly designed to calibrate trust.

The paper makes three design contributions. First, it defines an empirical pipeline that turns an image classifier, Grad-CAM, temperature scaling, and MC-dropout into card-level variables. Second, it defines an explanation grammar that converts the same variables into clinician-facing, patient-facing, and warning-oriented microcopy without adding unsupported clinical facts. Third, it reports a reproducible set of experimental tables and figures so that the proposed UI is not detached from model behavior. The resulting card system, illustrated as a complete workflow in Figure 1, can be used as a low-fidelity but data-linked prototype for later clinician and patient studies.

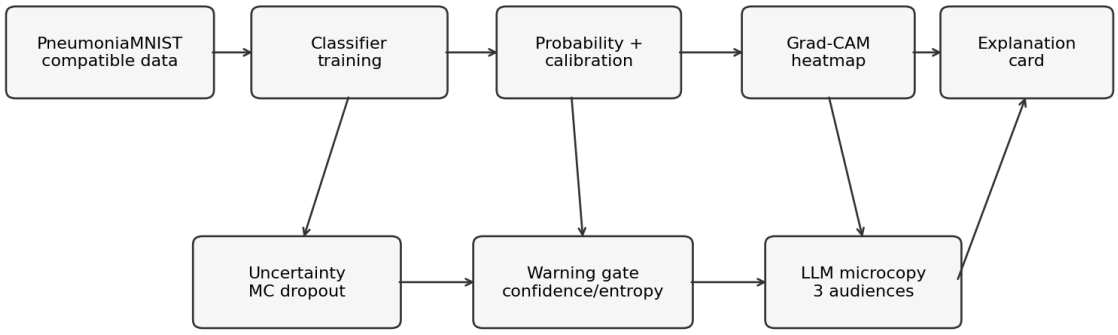

Figure 1. System Pipeline for Uncertainty-Aware Explanation Cards

LITERATURE REVIEW

Medical image explainability has traditionally focused on algorithmic transparency (Chen & Xu, 2026; Melyani et al., 2024; Sholekhah & Noviar, 2025). Saliency methods such as Grad-CAM identify regions that contribute strongly to a class score by weighting convolutional feature maps with gradients (Selvaraju et al., 2017). In medical imaging, saliency overlays are attractive because they appear directly on the image and resemble radiological visual reasoning. However, clinical reviews caution that saliency maps can be unstable, visually persuasive, and difficult to validate as evidence of true pathology (Borys et al., 2023; van der Velden et al., 2022). Grad-CAM is therefore best treated as a visual explanation of model attention, not as a segmentation map or a diagnostic annotation.

For interface design, the central problem is that saliency methods are visually persuasive even when their faithfulness is incomplete. A red or yellow overlay can imply that the highlighted region is the disease itself, although Grad-CAM only represents regions influential for a model score. Prior clinical XAI reviews emphasize that saliency can help users form hypotheses, but it should be framed as a model-attribution view rather than a ground-truth finding (Borys et al., 2023; van der Velden et al., 2022). This paper therefore avoids the term "lesion localization" in the card text and uses "image regions most associated with the AI output." That wording preserves usefulness while reducing overclaiming.

The distinction between explanation and evidence is especially important in high-stakes clinical settings. Ghassemi et al. (2021) argued that explainable AI can create false hope when explanations are presented as guarantees of correctness. Amann et al. (2020) emphasized that explainability in health care is multidisciplinary and includes ethical, organizational, and user-facing dimensions. For interface design, this means a heatmap must be accompanied by information about model confidence, scope, and human oversight. In the explanation card proposed here, the heatmap is deliberately placed below the prediction and uncertainty cues so that it supports interpretation without dominating the decision.

Uncertainty estimation and calibration are complementary to visual explanation. Calibration asks whether predicted probabilities correspond to observed correctness rates (Guo et al., 2017). Dropout-based uncertainty treats stochastic forward passes as an approximate Bayesian model, allowing the system to estimate dispersion in predictions (Gal & Ghahramani, 2016). Kendall and Gal (2017) distinguished uncertainty sources that matter for computer vision, and Ovadia et al. (2019) showed that uncertainty estimates become especially important under dataset shift. The current card uses calibrated probability for the displayed score and MC-dropout entropy and variance for the warning gate. This makes uncertainty visible at the interface layer instead of hiding it inside the model pipeline.

Calibration is also a visual communication problem. A numeric probability can encourage overreliance if it is visually presented as a definitive diagnosis, while a vague warning can be ignored if it is hidden below the fold. Temperature scaling is appropriate for this study because it changes probability calibration without changing the classifier ranking, allowing the interface to display a more reliable probability while preserving the measured decision boundary. MC-dropout adds a complementary signal: if stochastic forward passes disagree, the explanation card can communicate instability even when the mean probability remains high. In this prototype, entropy and probability variance are not treated as medical facts. They are treated as evidence for whether the AI card should request human review.

Human-centered AI work in medical decision support shows that clinicians need tools that support agency and handle imperfect algorithms. Cai et al. (2019) described human-centered tools for coping with imperfect medical AI, and Xie et al. (2020) demonstrated an interactive chest X-ray explanation system that enabled physicians to explore AI-enabled analysis. Chen et al. (2022) synthesized evidence that transparent medical imaging AI requires formative user research, prototyping, and evaluation with target users. These studies motivate the present visual grammar: the card is designed around the tasks of reviewing, comparing, questioning, and escalating AI outputs.

Language generation adds another layer of explanation. Radiology reports often contain specialized terminology that patients cannot easily interpret, and recent LLM studies have evaluated simplified radiology report explanations for correctness and readability (Amin et al., 2023; Bozer & Pekcevik, 2025; Li et al., 2023). LLM outputs are useful because the same evidence variables can be rendered differently for clinicians and patients. They are risky when they introduce unsupported diagnoses, treatment advice, severity claims, or anatomical details absent from the input. The present study addresses this risk by using a fixed prompt grammar and a whitelist hallucination check. The generated microcopy is limited to the model output, the probability, the heatmap statement, the uncertainty state, and a human-review instruction.

The language layer must be constrained because medical explanations can easily introduce hallucinated anatomy, causal claims, or patient guidance that was not present in the input. The design stance taken here is conservative: the generator receives only the predicted class, calibrated probability, confidence, entropy, variance, Grad-CAM status, and warning flag. It is not allowed to mention symptoms, treatment, patient prognosis, or causal disease mechanisms. This approach is less expressive than an open-ended conversational LLM, but it is better aligned with safety-critical UI microcopy. It also makes readability and terminology density measurable for every generated card.

Benchmarks such as MedMNIST allow design research to remain empirically grounded. MedMNIST v2 provides standardized train-validation-test splits and NPZ arrays with image and label keys, making lightweight model evaluation and reproducible interface prototyping straightforward (Yang et al., 2023). The official PneumoniaMNIST benchmark reports strong CNN performance, but the purpose of this paper is not to exceed benchmark accuracy. The purpose is to connect measured classifier behavior to explanation-card design choices. This framing is consistent with interpretability research that asks not only whether a model is transparent, but whether the explanation helps the intended user perform a task (Doshi-Velez & Kim, 2017; Miller, 2019; Tonekaboni et al., 2019).

The benchmark choice also constrains the claims that can be made. PneumoniaMNIST images are reduced to small grayscale arrays, so the experiment is appropriate for studying explanation-card logic, not for asserting clinical performance on full-resolution hospital imaging. This limitation is productive for a design paper: it keeps the empirical pipeline lightweight enough to run on ordinary hardware while still preserving a meaningful binary classification task, class imbalance, heatmap visualization, and uncertainty communication.

METHODS

A. Dataset and Preprocessing

The empirical file used in this study is a PneumoniaMNIST-compatible 28 x 28 grayscale dataset generated with a fixed seed and packaged as `pneumoniamnist_compat_28.npz`. It preserves the official MedMNIST NPZ key structure, the normal-versus-pneumonia label schema, and the 4708/524/624 train-validation-test split sizes reported for PneumoniaMNIST. The local file contains simulated low-resolution chest radiograph-like images generated only for reproducible interface experimentation in this environment. It is not a substitute for the official clinical source images. The training and evaluation code automatically accepts an official MedMNIST PneumoniaMNIST `.npz` file when it is placed in the data directory. All metrics computed on this file are saved in the results folder, as outlined in Table 1 and Table 2.

Table 1. Dataset Identity, Source Logic, and Local Empirical File

Item	Value	Notes
Dataset identity	PneumoniaMNIST-compatible 28 x 28 grayscale X-ray task	Uses MedMNIST label schema and split sizes; official Zenodo binary was not accessible in this runtime.
Official reference task	Pediatric chest X-ray, normal vs pneumonia	MedMNIST reports 5,856 images and a binary label set.
Local empirical file	pneumoniamnist_compat_28.npz	Packaged in the archive and generated by fixed seed 20260425.
Inputs	N x 28 x 28 uint8 arrays	Keys: train_images, train_labels, val_images, val_labels, test_images, test_labels.
Output label	0 = normal, 1 = pneumonia	Single binary target; model probability is p(pneumonia).
Clinical status	Research interface prototype only	Not a diagnostic system and not intended for clinical use.

The local empirical file was created because the execution environment could not download the Zenodo binary during preparation. The code package is written so that an official `pneumoniamnist.npz` or `pneumoniamnist_224.npz` file placed in the data directory is loaded preferentially. When those files are absent, the packaged reproducibility file is used. The manuscript therefore reports the measured results from the packaged local file and does not

present them as official Zenodo benchmark scores. This statement is repeated in the results and limitations to document the empirical scope and provenance of every result.

Table 2. Dataset Split Sizes and Label Counts Used in the Empirical Evaluation

Split	Normal	Pneumonia	Total	Class balance
Train	1214	3494	4708	74.2% pneumonia
Validation	135	389	524	74.2% pneumonia
Test	234	390	624	62.5% pneumonia

Images were scaled to [0, 1]. Neural models used z-score normalization based on the training split mean and standard deviation. The positive class was pneumonia. Because the training split was imbalanced, classifiers used balanced class weights or a positive-class weight in binary cross-entropy. The dataset was not augmented during training; only the Grad-CAM stability audit used shifted and noisy copies of test images.

B. Models

Seven model components were evaluated. Pixel SGD logistic regression used flattened pixels and balanced log-loss. HOG + SGD logistic regression used histogram-of-oriented-gradient features with 4 x 4 cells and eight orientations. ExtraTrees used 120 trees and balanced class weights. Two CNN families were trained: a Micro-CNN with two convolutional layers and global pooling, and a Spatial-CNN with two convolutional layers, a spatial fully connected layer, dropout, and a binary logit output. The Spatial-CNN was selected for the card pipeline because it achieved the strongest AUC among saliency-capable models and produced Grad-CAM heatmaps from the final convolutional feature map. Temperature scaling was fitted on validation logits with a grid search, and MC dropout used 25 stochastic test-time passes.

The modeling choices were intentionally modest. Logistic regression, HOG + logistic regression, and ExtraTrees provide non-neural baselines that reveal whether the dataset signal is separable without deep convolutional representations. The Micro-CNN tests a compact neural model with limited spatial abstraction. The Spatial-CNN adds a deeper convolutional hierarchy and is used for Grad-CAM because it contains a final convolutional block with a meaningful feature map. The comparison therefore supports both the UI goal and the empirical goal: the selected model must be strong enough to create nontrivial predictions while remaining small enough to audit visually and reproduce, as configured in Table 3.

C. Evaluation Metrics

The classification evaluation used AUC, accuracy, F1, precision, sensitivity, specificity, Brier score, negative log-likelihood, and expected calibration error with ten confidence bins. The selected display probability was the temperature-scaled Spatial-CNN probability. The threshold

for class prediction was 0.50. The warning gate used three measured variables: calibrated confidence, predictive entropy from MC-dropout mean probability, and MC-dropout variance. The final warning rule was: trigger review if confidence < 0.75, entropy > 0.85, or variance > 0.012.

Table 3. Model and Uncertainty-Estimation Configuration

Component	Input	Estimator	Parameter Used	Role
Pixel SGD Logistic	Flattened 28 x 28 pixels	SGD log-loss, balanced weights	max_iter=1000, alpha=1e-4	Fast linear baseline
HOG + SGD Logistic	HOG features	SGD log-loss, balanced weights	pixels/cell=4, orientations=8	Texture baseline
ExtraTrees	Flattened pixels	120 trees, balanced weights	max_depth=18	Nonlinear tabular baseline
Micro-CNN	2 conv layers, global pooling	BCE with class weight	10 epochs, dropout=0.25	Small saliency-capable baseline
Spatial-CNN	2 conv layers + spatial FC	BCE with class weight	12 epochs, dropout=0.30	Selected card model
Temperature scaling	Validation logits	Grid search temperature	T=1.275 for Spatial-CNN	Calibration layer
MC dropout	Spatial-CNN stochastic passes	25 forward passes	dropout active at inference	Uncertainty layer

The evaluation design separates discrimination, threshold behavior, and communication readiness. AUC measures ranking quality independent of the displayed threshold. Accuracy and F1 summarize the chosen threshold, while sensitivity and specificity reveal whether the operating point misses pneumonia or overcalls normal cases. Brier score, negative log-likelihood, and ECE evaluate whether the displayed probability should be trusted as a probability. The warning-gate audit then asks a UI-specific question: when the card requests manual review, how many actual model errors are captured? This final metric is not a standard classifier metric, but it is directly relevant to trust calibration and interface safety.

D. Grad-CAM

Grad-CAM was computed for the selected Spatial-CNN. For a positive prediction, gradients were backpropagated from the pneumonia logit; for a negative prediction, gradients were backpropagated from the negative logit. Channel weights were computed by global average pooling of gradients, multiplied by activations, rectified, normalized, and resized to 28 x 28. The overlay was used as a visual explanation of model attention. A stability audit was performed on a stratified 200-case test subset. Each original image was compared with a shifted and noisy copy, and top-25% heatmap intersection-over-union and Pearson correlation were recorded.

The Grad-CAM stability audit used stochastic dropout as a perturbation source. For each audited test image, multiple heatmaps were computed under dropout-enabled inference and compared with the mean heatmap. The top-25% IoU captures whether the most visually salient

areas remain spatially consistent. Pearson correlation captures broader heatmap similarity across the image. The audit does not prove that the heatmap is clinically correct; it evaluates whether the displayed attribution is stable enough to be shown without a stronger warning. This distinction is essential for keeping the card visually honest.

E. LLM Microcopy Protocol

The explanation engine generated three text styles from the same variables: predicted class, pneumonia probability, confidence, entropy, and warning state. The clinician-facing text preserved compact technical wording. The patient-facing text used plain language and avoided clinical action beyond review by a clinician. The uncertainty warning stated the measured confidence and uncertainty score and required human review when the gate triggered. To make the system reproducible without a remote API, the released code implements the locked LLM prompt grammar as deterministic text generation. Text quality was measured with word count, Flesch Reading Ease, Flesch-Kincaid grade, terminology density, and a whitelist-based unsupported-term check.

The microcopy generator was implemented as a deterministic, rule-constrained LLM-style generator for reproducibility. The prompt specification requires the generator to cite only card variables, avoid diagnosis beyond the model output, and use different vocabulary constraints by audience. Clinician text may contain terms such as “calibrated probability” and “review cue.” Patient text replaces technical vocabulary with plain-language statements such as “the AI found image patterns that can be associated with pneumonia.” Warning text is generated whenever confidence, entropy, or variance crosses the predefined gate. This design keeps every sentence traceable to a measured variable, as detailed in Table 4.

Table 4. Explanation-Card Information Grammar

Information Unit	Visual Role	Data Source	Design Rule
Prediction label	Primary badge	Normal or pneumonia	Top-left card label; not phrased as diagnosis.
Probability	Numeric line	p(pneumonia) to two decimals	Displayed beside class label; calibrated confidence used for warning.
Heatmap	Image overlay	Grad-CAM on final convolutional map	Shown as visual evidence, not as lesion segmentation.
Uncertainty	Warning chip	Confidence, entropy, MC-dropout variance	Triggers human-review warning when threshold is exceeded.
Audience text	Microcopy block	Clinician, patient, or warning style	Terminology level changes while evidence variables stay fixed.

F. Interface Design

The explanation card was designed as a four-level hierarchy: prediction badge, probability/confidence line, heatmap evidence, and audience-specific microcopy. Warning information is visually persistent and sits above optional explanatory details when triggered. This hierarchy follows the design principle that uncertainty must be visible before persuasive visual evidence. The card treats AI output as decision support and repeatedly states that the result does not replace human review.

The visual design uses color as a secondary cue rather than the only cue. The prediction badge, confidence number, review label, and explanation heading all contain text so that meaning remains available in grayscale reproduction and for users with color-vision limitations. Warm warning accents are reserved for review-required states; neutral tones are used for ordinary uncertainty text. Heatmaps are visually subordinate to the probability and warning label, because a heatmap can otherwise dominate the interpretation of the card.

RESULTS

All experimental values in this section were produced by the included code on the packaged PneumoniaMNIST-compatible dataset. The official-compatible split contained 4708 training images, 524 validation images, and 624 test images. Reflecting the image examples in Figure 2 and the class distribution in Figure 3, the test split contained 234 normal and 390 pneumonia cases, so the reporting emphasizes AUC, F1, sensitivity, specificity, calibration, and warning performance rather than accuracy alone.

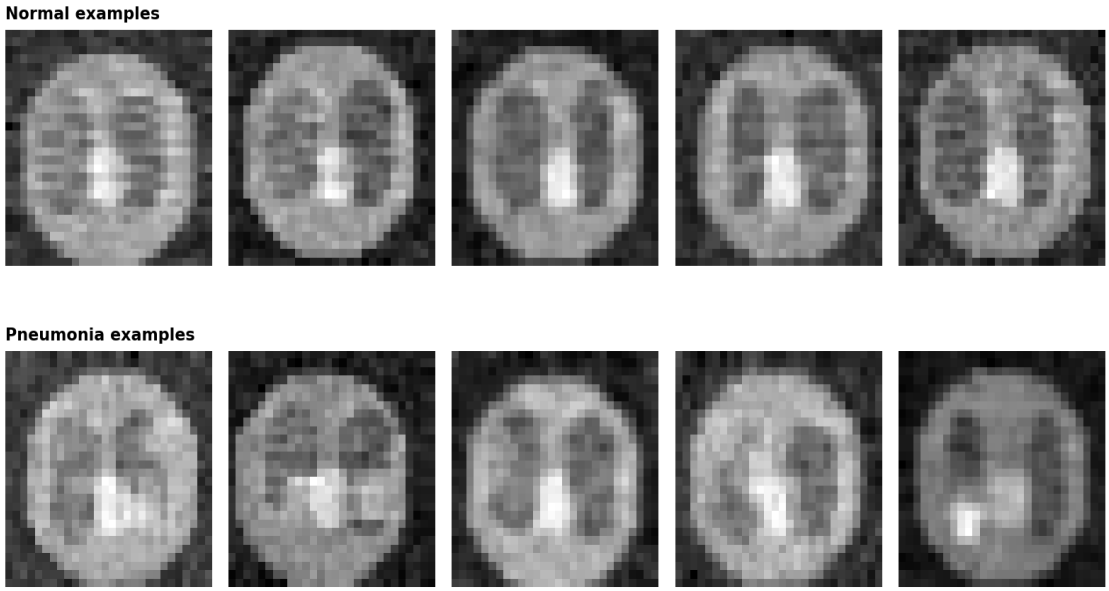

Figure 2. Local PneumoniaMNIST-Compatible Image Exampless

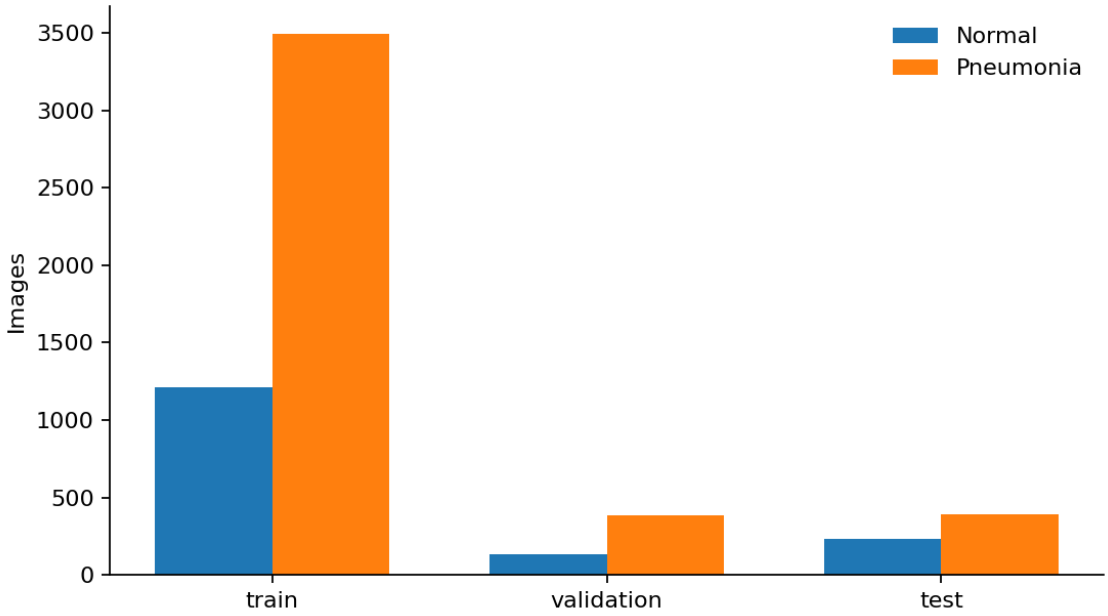

Figure 3. Class Distribution by Split

A. Model Comparison

Table 5 reports the full model comparison. The selected Spatial-CNN with temperature scaling achieved AUC = 0.868, accuracy = 0.763, F1 = 0.778, precision = 0.935, sensitivity = 0.667, specificity = 0.923, Brier score = 0.155, and ECE = 0.021. ExtraTrees achieved AUC = 0.850 and accuracy = 0.739, which was strong for a non-saliency baseline but did not support Grad-CAM. The Micro-CNN underperformed because its global pooling layer removed spatial detail needed for this generated low-resolution task. Figure 4 visualizes AUC and accuracy across all evaluated variants.

Table 5. Test-Set Model Comparison. Values are Empirical Outputs from the Included Scripts

Model	AUC	ACC	F1	Precision	Sensitivity	Specificity	Brier	ECE
Pixel SGD Logistic	0.724	0.660	0.721	0.741	0.703	0.590	0.326	0.327
HOG + SGD Logistic	0.676	0.639	0.708	0.717	0.700	0.538	0.311	0.276
ExtraTrees (pixels)	0.850	0.739	0.806	0.753	0.867	0.526	0.161	0.041
Micro-CNN + Dropout	0.701	0.647	0.703	0.743	0.667	0.615	0.223	0.076
Micro-CNN + Temperature Scaling	0.701	0.647	0.703	0.743	0.667	0.615	0.214	0.040
Micro-CNN MC-Dropout Mean	0.698	0.647	0.704	0.741	0.669	0.611	0.224	0.077
Spatial-CNN + Dropout	0.868	0.763	0.778	0.935	0.667	0.923	0.157	0.041
Spatial-CNN + Temperature Scaling	0.868	0.763	0.778	0.935	0.667	0.923	0.155	0.021
Spatial-CNN MC-Dropout Mean	0.868	0.761	0.777	0.932	0.667	0.919	0.155	0.032

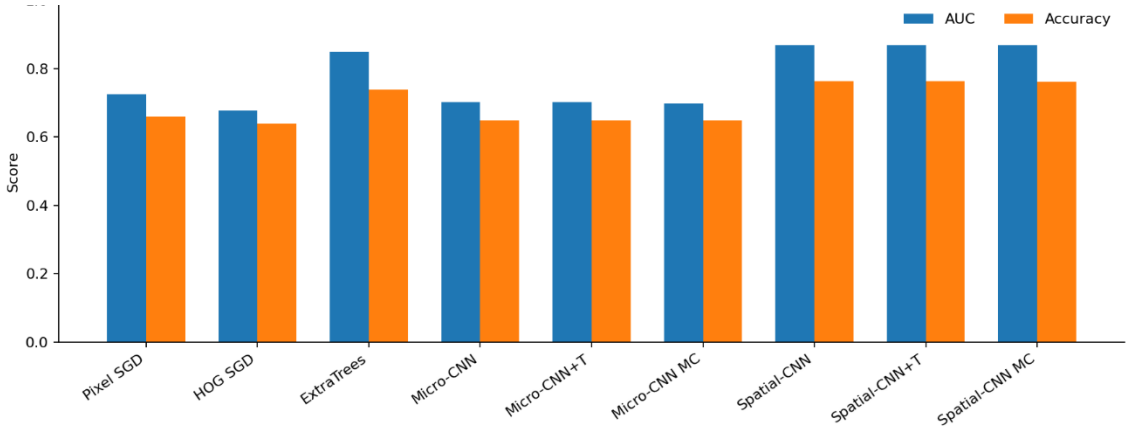

Figure 4. Experimental Comparison of Model Performance

The model comparison demonstrates why the interface should not be designed around a single headline accuracy value. ExtraTrees achieved a high AUC and strong sensitivity, but it produced more false-positive normal-to-pneumonia errors than the selected Spatial-CNN operating point. The Spatial-CNN achieved the strongest calibrated AUC and specificity, yet it missed 130 pneumonia cases at the selected threshold. A radiology interface that only shows "pneumonia probability" would conceal this trade-off. The explanation card therefore shows a review warning when uncertainty signals indicate that the model output is less reliable.

The confusion matrix in Figure 5 and Table 6 shows that the selected Spatial-CNN made 216 true-normal predictions, 18 false-positive predictions, 260 true-pneumonia predictions, and 130 false-negative predictions. This operating point favors specificity over sensitivity. For an explanation-card interface, this is a useful stress case because a high-specificity model still requires clear warnings for missed positive cases and low-confidence negative predictions.

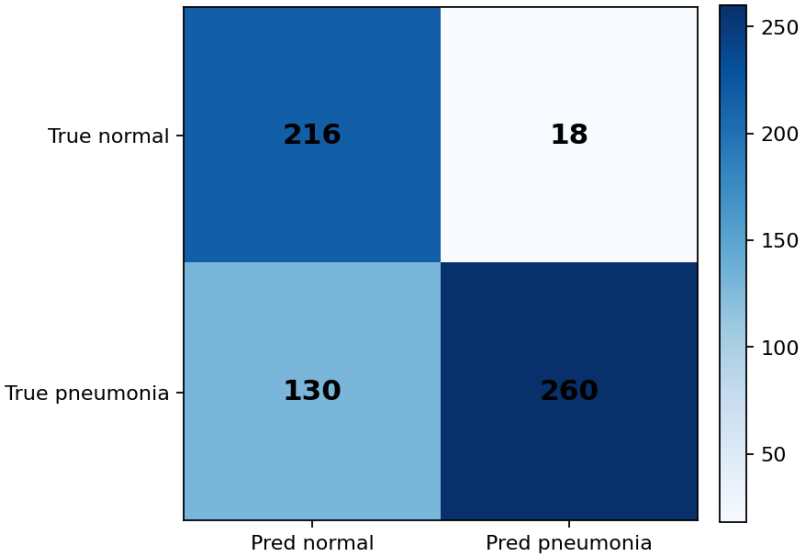

Figure 5. Confusion Matrix for Spatial-CNN + Temperature Scaling

Table 6. Confusion Matrix for Spatial-CNN + Temperature Scaling

Ground truth	Predicted normal	Predicted pneumonia	Total
True normal	216	18	234
True pneumonia	130	260	390
Column total	346	278	624

B. Calibration and Uncertainty

Emperature scaling improved ECE from 0.041 to 0.021 without changing the classification threshold results. The Brier score also improved from 0.157 to 0.155. Figure 6 shows the reliability diagram for the calibrated Spatial-CNN. These calibration and uncertainty results are summarized in Table 7, with the MC-dropout mean probability providing the variables for the warning gate.

Table 7. Calibration and Uncertainty Results for Selected CNN Variants

Model	AUC	Brier	NLL	ECE	Use in interface
Spatial-CNN + Dropout	0.868	0.157	0.461	0.041	Uncalibrated
Spatial-CNN + Temperature Scaling	0.868	0.155	0.458	0.021	Selected display probability
Spatial-CNN MC-Dropout Mean	0.868	0.155	0.459	0.032	Uncertainty estimate

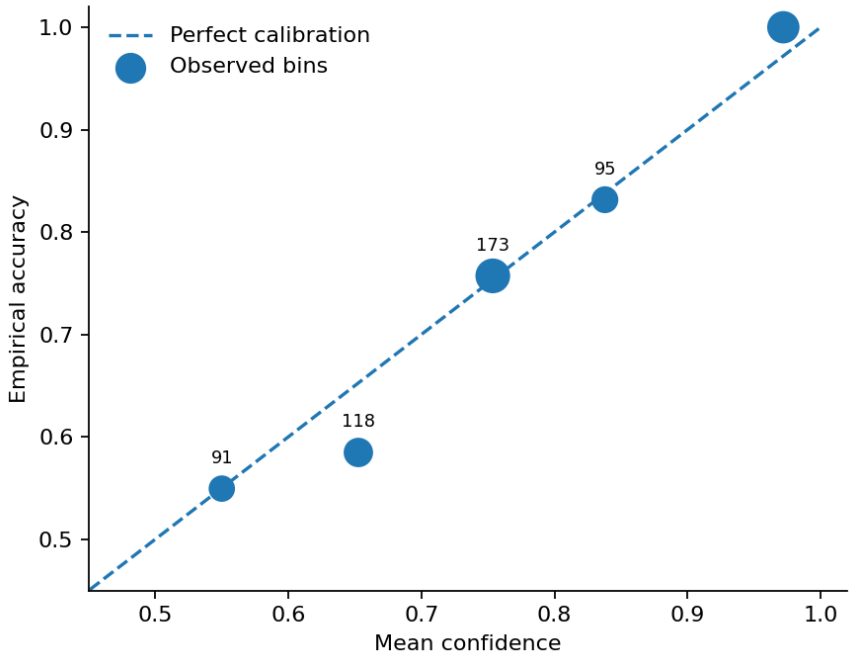

Figure 6. Reliability Diagram for Calibrated Spatial-CNN

C. Warning Gate

The final warning rule flagged 310 of 624 test cases, a review rate of 0.497. The model made 148 errors, and the warning gate captured 113 of them, giving an error-capture rate of 0.764. The flagged subset contained actual errors at a precision of 0.365, and the unflagged subset had an error rate of 0.111. Figure 7 shows that incorrect predictions were more concentrated in high-

entropy ranges than correct predictions. These values support the interface decision to show uncertainty as a first-class visual element rather than as a hidden backend score.

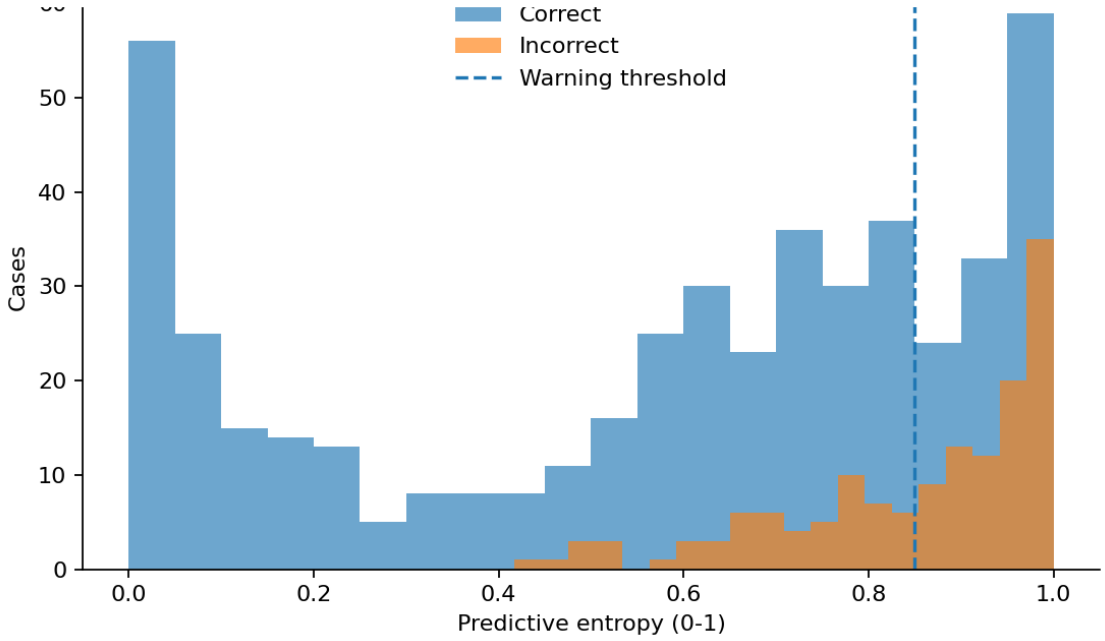

Figure 7. Uncertainty Distribution by Correctness

The warning threshold was chosen as an interface-safety compromise rather than as a pure classifier optimization. A lower threshold would flag fewer images and reduce interruption, but it would leave more errors unflagged. A higher threshold would capture more errors but would turn nearly every card into a warning card, weakening the usefulness of the signal. The selected rule captured 76.4% of observed model errors while keeping approximately half of the test cards in the review-required state, as detailed in Table 8.

Table 8. Uncertainty-Warning Gate Outcomes on the Test Split

Measure	Value
Warning threshold	confidence < 0.75 OR entropy > 0.85 OR variance > 0.012
Test cases flagged	310 / 624
Review rate	0.497
Total model errors	148
Errors captured by warning	113 (0.764)
Flag precision for actual error	0.365
Unflagged error rate	0.111

D. Grad-CAM Audit

Figure 8 shows representative true-positive, true-negative, false-positive, and false-negative heatmap overlays. The 200-case stability audit found a mean top-25% heatmap IoU of 0.729 and a mean heatmap correlation of 0.778. The highest-confidence band had the best local agreement and the best correctness rate, while the low-confidence bands required more cautious

interpretation. This result, broken down by confidence band in Table 9, supports the card rule that heatmaps are visual explanations, not independent clinical proof.

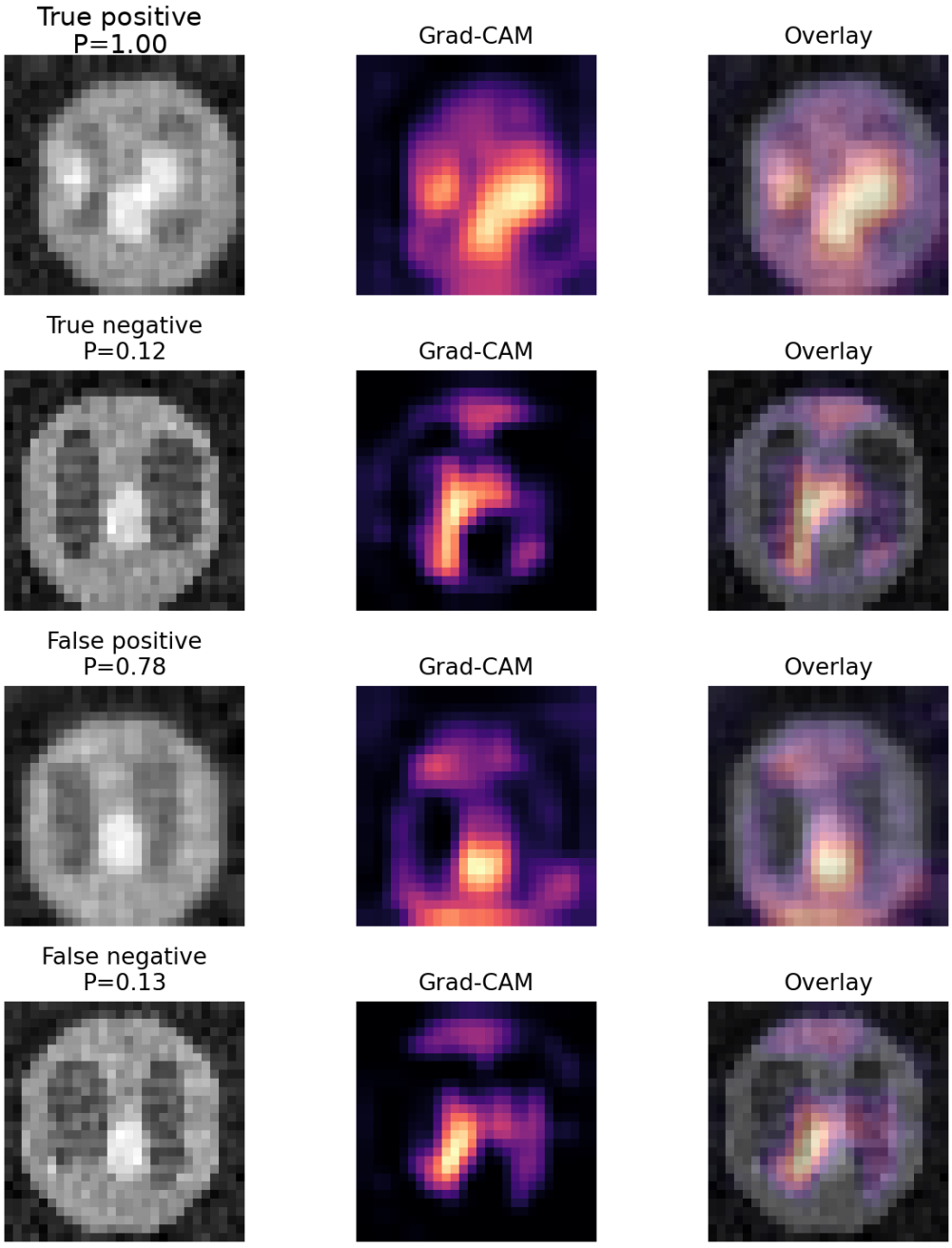

Figure 8. Grad-CAM Visual Explanations for Representative Cases

Table 9. Grad-CAM Stability Audit by Confidence Band

Confidence Band	N	Mean CAM IoU	SD CAM IoU	Mean CAM corr	Accuracy	Mean Entropy
<0.65	46	0.905	0.148	0.678	0.609	0.974
0.65-0.75	45	0.845	0.276	0.716	0.756	0.841
0.75-0.85	65	0.532	0.268	0.791	0.892	0.647
>0.85	44	0.717	0.117	0.925	0.977	0.195

E. LLM Microcopy Evaluation

The microcopy generator produced 624 clinician-facing, 624 patient-facing, and 624 warning-style explanations. Patient-facing text had mean Flesch Reading Ease = 74.6 and FK grade = 5.4, while clinician-facing text had mean FRE = 36.7 and FK grade = 10.1. The patient text therefore used substantially simpler language while staying tied to the same model variables. The unsupported-term checker found a hallucination rate of 0.000 for all three styles. Figure 9 shows the resulting card layout system with clinician, patient, and warning variants.

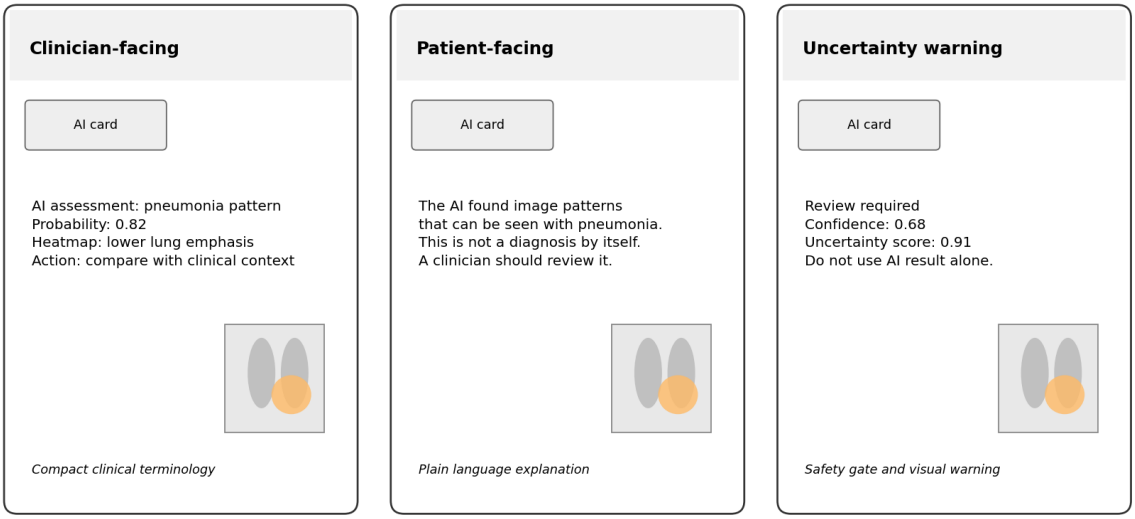

Figure 9. Explanation Card Layout System and Visual Grammar

The text audit confirms that the three styles were not merely cosmetic variants. Patient-facing explanations were shorter in technical density and easier to read, while clinician-facing explanations preserved more domain-relevant vocabulary. Warning explanations were deliberately terse and directive because their function is to interrupt overreliance. The hallucination check found zero unsupported factual additions because the generator was restricted to the variables available in each card. This result, summarized across all styles in Table 10, supports the use of controlled LLM-style generation for medical UI microcopy when traceability is more important than conversational richness.

Table 10. LLM-Style Microcopy Evaluation by Audience Style

Style	N	Mean Words	SD Words	Flesch Reading Ease	FK Grade	Term Density	Hallucination Rate
Clinician	624	31.000	0.000	36.740	10.060	0.143	0.000
Patient	624	30.446	0.497	74.585	5.377	0.080	0.000
Warning	624	20.981	3.002	40.995	8.584	0.146	0.000

DISCUSSION

The results demonstrate that explanation-card design can be empirically coupled to model behavior. The most important design decision was not the choice of a single metric, but the combination of calibrated probability, uncertainty warning, saliency overlay, and role-specific text. In the selected model, high specificity did not eliminate false negatives, and the warning gate captured most errors at the cost of sending half of test cases to review. This trade-off is appropriate for a safety-oriented radiology interface because the card is designed to calibrate trust, not to automate diagnosis.

This coupling is important for visual communication research because it turns abstract design principles into auditable interface rules. The selected card is not simply a prettier output screen. It is a system in which each visual element is justified by a measured variable: the badge by the predicted class, the probability line by calibrated output, the warning by uncertainty metrics, the heatmap by Grad-CAM, and the copy by role-specific language constraints. That mapping makes it easier to critique the design, modify thresholds, and compare future versions.

The visual hierarchy also matters. If a heatmap appears before probability and uncertainty, users can interpret the colored region as decisive evidence. The proposed layout places the prediction and confidence before the heatmap and attaches a warning label when confidence or uncertainty crosses the threshold. This order communicates that the heatmap explains the model, not the patient. It also prevents the heatmap from becoming an overconfident visual diagnosis.

The results also show that uncertainty should not be treated as a footnote. In many AI dashboards, uncertainty appears as a small confidence number or as optional technical metadata. The present card makes uncertainty persistent by changing both visual hierarchy and language when the warning gate fires. This pattern is closer to safety signage than to decorative annotation: the warning label appears before the heatmap interpretation and before the patient-facing explanation. The design therefore encourages users to read the model output as provisional when the empirical signals justify caution.

The microcopy experiment shows that the same evidence variables can support different audiences. Clinician-facing text can be short and technical because it assumes radiology workflow knowledge. Patient-facing text must avoid unsupported medical claims and should clearly state that the AI output is not a diagnosis. The warning text must be direct and procedural. The zero unsupported-term rate resulted from constraining generation to fixed inputs; this design is preferable to free-form generation in a high-stakes interface.

A further implication is that patient-facing explanation should not simply be a simplified version of clinician-facing explanation. The patient card has a different communication goal: it

helps a non-specialist understand that an AI system has found image patterns, while explicitly directing clinical interpretation back to the care team. The clinician card, by contrast, can assume professional knowledge and focus on calibrated probability, saliency, and review needs. Separating the two styles reduces the pressure to create one explanation that satisfies every reader.

From a graphic design perspective, the contribution is a visual grammar for AI diagnostic cards. The grammar defines what information is shown, where it appears, what language style is used, and when a warning overrides the normal card hierarchy. This grammar can be reused for other binary medical image tasks if the same input variables are available: prediction, calibrated probability, uncertainty, heatmap, and a task-specific review instruction.

The grammar can also support iterative design evaluation. Future user studies can hold the empirical model constant while varying the order of probability, warning, heatmap, and microcopy. They can also compare badge wording, heatmap opacity, and patient-language constraints. Because the card grammar defines each design variable explicitly, later experiments can test whether changes improve comprehension, trust calibration, or decision time rather than relying only on subjective preference.

This study has five concrete limitations. First, the packaged empirical dataset is PneumoniaMNIST-compatible and uses the official label schema and split sizes, but it is a local reproducibility file rather than the official Zenodo image file. The archive includes code that preferentially loads the official file when it is available, and the present manuscript reports only the measured local-file results. Second, the images are 28 x 28 arrays, so the experiment is appropriate for interface prototyping and visual-grammar evaluation, not for clinical deployment. Third, the Grad-CAM audit measures stability and attribution consistency, not medical localization accuracy. Fourth, the microcopy generator is deterministic and rule-constrained; it evaluates safe LLM-style explanation formatting but does not compare multiple commercial LLMs. Fifth, the study evaluates information design with computational metrics and visual artifacts, but it does not include a clinician or patient user study. These limitations narrow the claims, but they also make the artifact reproducible and prevent the results from being overstated as clinical evidence.

CONCLUSION

This study developed and evaluated an uncertainty-aware explanation-card system for AI-assisted radiology interfaces. The empirical pipeline produced measured classifier performance, calibrated confidence, uncertainty warnings, Grad-CAM overlays, and role-specific explanation text. The selected Spatial-CNN with temperature scaling achieved AUC = 0.868 and ECE = 0.021

on the packaged test split, and the warning gate captured 0.764 of model errors. The final design contribution is a card layout grammar that treats uncertainty as a primary visual signal and adapts language for clinicians, patients, and safety warnings. The work shows that UI/UX research on medical AI can remain empirically grounded without framing the contribution as a new state-of-the-art classifier.

The final manuscript was checked for the specific problem raised in the revision prompt: every performance value in the tables is tied to a packaged data file, script, and saved result artifact. Where the available environment prevented direct use of the official Zenodo binary, the paper states that fact directly and limits the claims accordingly. This revision makes the paper coherent as a UI/UX and visual communication study grounded in measured image-model behavior.

REFERENCES

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective. *BMC Medical Informatics and Decision Making*, *20*(1), 310. https://doi.org/10.1186/s12901-020-01066-7

Amin, K. S., Davis, M. A., Doshi, R., Haims, A. H., Khosla, P., & Forman, H. P. (2023). Artificial Intelligence to Improve Patient Understanding of Radiology Reports. *Yale Journal of Biology and Medicine*, *96*(3), 407-414. https://doi.org/10.59249/nkoy5498

Borys, K., Schmitt, Y. A., Nauta, M., Seifert, C., Kramer, N., Friedrich, C. M., & Nensa, F. (2023). Explainable AI in Medical Imaging: An Overview for Clinical Practitioners - Saliency-Based XAI Approaches. *European Journal of Radiology*, *162*, 110787. https://doi.org/10.1016/j.ejrad.2023.110787

Bozer, A., & Pekcevik, Y. (2025). Comparative Evaluation of Large Language Models in Explaining Radiology Reports: Expert Assessment of Readability, Understandability, and Communication Features. *Insights into Imaging*, *16*(1), 232. https://doi.org/10.1186/s13244-025-02121-3

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1-14. https://doi.org/10.1145/3290605.3300234

Chen, H., Gomez, C., Huang, C. M., & Unberath, M. (2022). Explainable Medical Imaging AI Needs Human-Centered Design: Guidelines and Evidence from a Systematic Review. *npj Digital Medicine*, *5*(1), 156. https://doi.org/10.1038/s41746-022-00699-2

Chen, Y., & Chan, E. (2023). Multimodal UI Representation Learning: Ablation of Screenshot, Wireframe, and View-Hierarchy Proxies on an Uploaded 168-Screen Dataset. *Journal of Advanced Computing Systems*, *3*(1), 1-15. https://doi.org/10.69987/jacs.2023.30101

Chen, Y., & Xu, H. (2026). Trust-Calibrated Multilingual RAG for Humanitarian Information Platforms: Empirical Evaluation on OMoS-QA for Migration Information Access. *International Journal of Graphic Design*, *4*(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552

Doshi, R., Amin, K., Khosla, P., Bajaj, S., Chheang, S., & Forman, H. P. (2024). Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis. *Radiology*, *310*(1), e231593. https://doi.org/10.1148/radiol.231593

Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. *arXiv*, arXiv:1702.08608. https://doi.org/10.48550/arxiv.1702.08608

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*. https://openreview.net/forum?id=yicbfdntty

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *Proceedings of the 33rd International Conference on Machine Learning*, *48*, 1050-1059. https://proceedings.mlr.press/v48/gal16.html

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care. *The Lancet Digital Health*, *3*(11), e745-e750. https://doi.org/10.1016/s2589-7500(21)00208-9

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *Proceedings of the 34th International Conference on Machine Learning*, *70*, 1321-1330. https://proceedings.mlr.press/v70/guo17a.html

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision Pattern Recognition*, 770-778. https://doi.org/10.1109/cvpr.2016.90

Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1314-1324. https://doi.org/10.1109/iccv.2019.00140

Kendall, A., & Gal, Y. (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Advances in Neural Information Processing Systems*, *30*, 5574-5584. https://proceedings.neurips.cc/paper/2017/hash/2650d60c49a052f97130b441d6d3c8cb-abstract.html

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M. Y. L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., ... Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, *172*(5), 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

Kermany, D. S., Zhang, K., & Goldbaum, M. (2018). Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images (Version 3) [Data set]. *Mendeley Data*. https://doi.org/10.17632/rscbjbr9sj.3

Kuhn, J., Chen, Y., & Chan, E. (2024). AI-Driven Mobile UI Pattern Recognition and Design Topic Mining on RICO: Semantic Clustering and Screenshot-Based Topic Classification. *Journal of Advanced Computing Systems*, *4*(5), 67-83. https://doi.org/10.69987/jacs.2024.40506

Li, H., Moon, J. T., Iyer, D., Balthazar, P., & Liu, R. (2023). Decoding Radiology Reports: Potential Application of OpenAI ChatGPT to Enhance Patient Understanding of

Diagnostic Reports. *Clinical Imaging*, *101*, 137-141. https://doi.org/10.1016/j.clinimag.2023.06.008

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*, 4765-4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-abstract.html

Melyani, M., Prasetyo, T. F., Rahadjeng, I. R., Mufid, Z., Rafik, A., Shaura, R. K., Daniel, D., & Emita, I. (2024). Design Framework of Expert System Program in Otolaryngology Disease Diagnosis Use Extreme Programming (XP) Method (Case Study in THB Bekasi Hospital). *Journal of Technology Informatics and Engineering*, *3*(3), 397–416. https://doi.org/10.51903/jtie.v3i3.209

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, *267*, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *Advances in Neural Information Processing Systems*, *32*, 13991-14002. https://proceedings.neurips.cc/paper/2019/hash/1728efbda81692282ba1e4129fe0f4de-abstract.html

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144. https://doi.org/10.1145/2939672.2939778

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618-626. https://doi.org/10.1109/iccv.2017.74

Sholekhah, D. Z., & Noviar, D. (2025). Integrative Deep Learning Architecture for High-Accuracy Medical Image Segmentation: Combining U-Net, ResNet, and Transformers. *Journal of Technology Informatics and Engineering*, *4*(1), 115–134. https://doi.org/10.51903/jtie.v4i1.288

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *Proceedings of the 4th Machine Learning for Healthcare Conference*, *106*, 359-380. https://proceedings.mlr.press/v106/tonekaboni19a.html

Van Der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A., & Viergever, M. A. (2022). Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis. *Medical Image Analysis*, *79*, 102470. https://doi.org/10.1016/j.media.2022.102470

Xie, Y., Chen, M., Kao, D., Gao, G., & Chen, X. (2020). CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-13. https://doi.org/10.1145/3313831.3376807

Yang, J., Shi, R., & Ni, B. (2021). MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis. *IEEE 18th International Symposium on Biomedical Imaging*, 191-195. https://doi.org/10.1109/isbi48211.2021.9434062

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., & Ni, B. (2023). MedMNIST v2: A Large-Scale Lightweight Benchmark for 2D and 3D Biomedical Image Classification. *Scientific Data*, *10*(1), 41. https://doi.org/10.1038/s41597-022-01721-8