International Journal of Graphic Design (IJGD)
Volume 3 No. 2 October 2025
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Page 365-380
DOI: https://doi.org/10.51903/ijgd.v3i2.3701
Available online at: https://journal.stekom.ac.id/index.php/ijgd

Uncertainty-Aware Breast Ultrasound Explanation Cards: A Visual Communication Framework for Image-Based AI Diagnostic Support Using BreastMNIST_224

Tong Ye[1], Xiaohan Chang*[2], Eric Zhong[3]
[1]Computer Science, Northeastern University, CA, USA
[2] Computer Science, University of Connecticut, CT, USA
[3] Computer Science, USC, CA, USA
Email Address: akiraye1999@gmail.com

Abstract. *AI-assisted diagnostic interfaces should communicate more than a class label. They need to show the predicted risk, the uncertainty around that risk, the visual evidence that influenced the model, the limits of the evidence, and the appropriate next action. This paper presents an uncertainty-aware explanation-card framework for breast ultrasound decision-support screens. The empirical study was conducted on BreastMNIST_224, the 224 x 224 MedMNIST+ breast ultrasound benchmark with official train, validation, and test splits of 546, 78, and 156 images. The positive class was defined as malignant. Five image classifiers were trained on downsampled image grids, and the selected card model was a Platt-probability RBF SVM. On the official test split, the selected model achieved AUROC = 0.867 and AUPRC = 0.728. A validation-selected operating threshold of 0.254 gave accuracy = 0.769, sensitivity = 0.833, specificity = 0.746, Brier score = 0.125, and ECE = 0.068. The explanation card pairs malignant-risk probability with risk tier, uncertainty band, occlusion-sensitivity heatmap evidence, a limitation statement, and a review cue. In the held-out test set, the conservative Low-risk tier contained six cases and no malignant cases; all seven false negatives occurred in the Review tier rather than in Low risk. These findings support a prototype-level visual communication framework in which image evidence is shown together with uncertainty and safeguards, while diagnostic authority remains with the clinician.*

Keywords *: Breast Ultrasound, Diagnostic Interface, Explainable Artificial Intelligence, Explanation Cards, Uncertainty Visualization.*

INTRODUCTION

Medical AI interfaces are not neutral containers for model output. In diagnostic settings, visual hierarchy, probability language, color, spatial emphasis, and explanatory order can shape whether a user treats an algorithm as decision support or as an implied diagnosis. Explainable artificial intelligence is therefore also a communication problem: an interface must show how a prediction was formed, what uncertainty accompanies it, and what action is appropriate when the model is confident, ambiguous, or wrong (Gunning et al., 2019; Miller, 2019; Tonekaboni et al., 2019).

This paper studies breast cancer decision-support communication through explanation cards. An explanation card is a compact UI component that summarizes the model output in a stable information architecture: malignant-risk probability, uncertainty, visual evidence, limitation, and next action. The goal is not to claim clinical readiness or to maximize algorithmic novelty. The goal is to test whether a reproducible image-based model output can be translated into a cautious visual artifact that supports appropriate review rather than overtrust.

Received: July 2025; Revised: August 2025; Accepted: September 2025; Published: October 2025
*Corresponding author, akiraye1999@gmail.com

The empirical study uses BreastMNIST_224, an image-based breast ultrasound dataset from MedMNIST+. This addresses the visual communication problem directly because the evidence area of the card can display an ultrasound thumbnail and an occlusion-sensitivity heatmap rather than only numeric feature contributions. Figure 1 summarizes the workflow linking the ultrasound image, classifier probability, calibration and uncertainty bands, visual evidence, and card-level follow-up guidance.

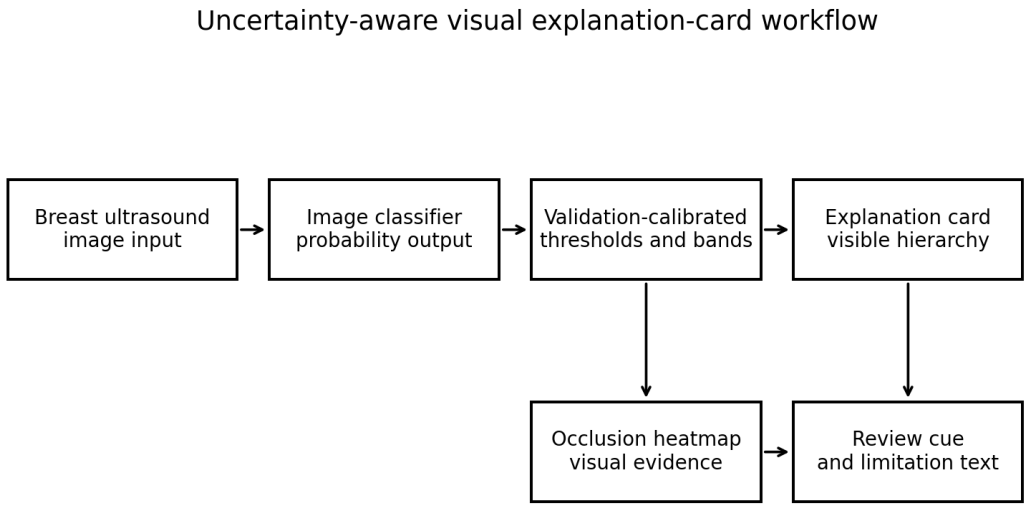

Figure 1. Visual communication workflow linking image input, probability output, uncertainty bands, heatmap evidence, and card-level review guidance

The research contribution is threefold. First, the paper reports a fully executed image-based benchmark on the official BreastMNIST_224 split, including discrimination, clinical error, and probability-quality metrics. Second, it implements a deterministic explanation-card generator that converts model output into a consistent visual communication structure. Third, it evaluates the card through computational UI checks, including risk-tier behavior, uncertainty-band behavior, representative cases, and information-coverage audit. The audit is intentionally limited: it verifies that the card contains the required communication elements, but it does not claim that clinicians understand, trust, or use the card better without a user study.

The intended audience is a visual communication researcher, UI/UX evaluator, or designer building a diagnostic-support prototype. For that audience, the relevant question is not only whether a classifier performs well, but whether the interface prevents a probability from being read as a diagnosis. The proposed visual hierarchy therefore places risk and uncertainty above the evidence thumbnail and places the limitation and next action directly inside the card.

LITERATURE REVIEW

Breast cancer AI studies often emphasize classification performance, but interface-centered deployment requires a broader theory of explanation. BreastMNIST is useful for this purpose because it standardizes breast ultrasound images into a lightweight binary classification task with malignant versus normal/benign labels (Yang et al., 2021, 2023). Because the input is an image, explanation cards can include spatial evidence such as heatmaps, overlays, or paired original-and-evidence tiles rather than feature-only lists.

XAI literature distinguishes local explanations, global explanations, and explanation formats designed for human decision-making. Local explanation methods such as LIME and SHAP were designed to explain individual predictions in model-agnostic or attribution terms (Lundberg & Lee, 2017; Ribeiro et al., 2016). Vision-oriented techniques such as saliency maps, occlusion sensitivity, and Grad-CAM provide spatial cues that can be overlaid on images (Simonyan et al., 2014; Zeiler & Fergus, 2014; Selvaraju et al., 2017). These visualizations are useful, but they should not be presented as clinical proof. A heatmap can indicate model sensitivity without guaranteeing that the highlighted region corresponds to a lesion boundary.

Calibration is central to uncertainty-aware interface design (Kuhn et al., 2024). A model can rank cases well while assigning probabilities that are too high or too low. Reliability diagrams, Brier score, negative log-likelihood, and expected calibration error capture this difference between discrimination and probability quality (Guo et al., 2017; Niculescu-Mizil & Caruana, 2005). In a diagnostic interface, calibration affects triage language because a 0.80 malignant-risk display communicates a different level of concern than a 0.30 display, even when both might require review.

Human-centered AI (Chen & Chan, 2023) research emphasizes appropriate reliance. Clinicians and trainees need explanations that match workflow, use domain-relevant cues, and clearly communicate uncertainty and system limits (Amershi et al., 2019; Cai et al., 2019; Tonekaboni et al., 2019). Medical AI deployment studies similarly warn against treating algorithmic output as a finished clinical product without validation, integration, and oversight (Kelly et al., 2019). This paper follows that caution by narrowing its claim to a prototype-level communication framework evaluated through computational checks rather than human-subject outcomes.

The visual communication challenge is that explanation can become persuasion. A polished heatmap or confidence badge may increase apparent authority even when a case is ambiguous. The proposed card therefore displays uncertainty and limitations in the same compact object as the risk score and heatmap. False negatives, false positives, and borderline probabilities are treated as important interface cases rather than hidden error cases.

METHODS

All experiments used the BreastMNIST_224 NPZ file in the MedMNIST format. The file contains train_images, train_labels, val_images, val_labels, test_images, and test_labels. Images are 224 x 224 grayscale breast ultrasound images. The official label mapping is 0 for malignant and 1 for normal/benign; for evaluation, malignant was treated as the positive class. Table 1 defines the empirical dataset role, and Table 2 reports the split and class counts used in the experiments. Figure 2 shows representative ultrasound images from the official test split.

Table 1. Dataset source alignment and empirical role

Dataset	Data form	Task and labels	Use in this study
BreastMNIST_224	224 x 224 grayscale breast ultrasound images	Binary classification; malignant versus normal/benign	Primary empirical dataset for image classification, probability display, and occlusion-heatmap explanation cards
Official validation split	78 labeled ultrasound images	Same binary label mapping	Operating-threshold selection and risk-band checks
Official test split	156 labeled ultrasound images	Same binary label mapping	Held-out reporting for discrimination, clinical error, calibration, risk tiers, uncertainty bands, and representative cards

Table 2. Official BreastMNIST_224 split used for empirical evaluation

Split	N	Malignant	Normal/benign	Malignant rate
train	546	147	399	0.269
validation	78	21	57	0.269
test	156	42	114	0.269

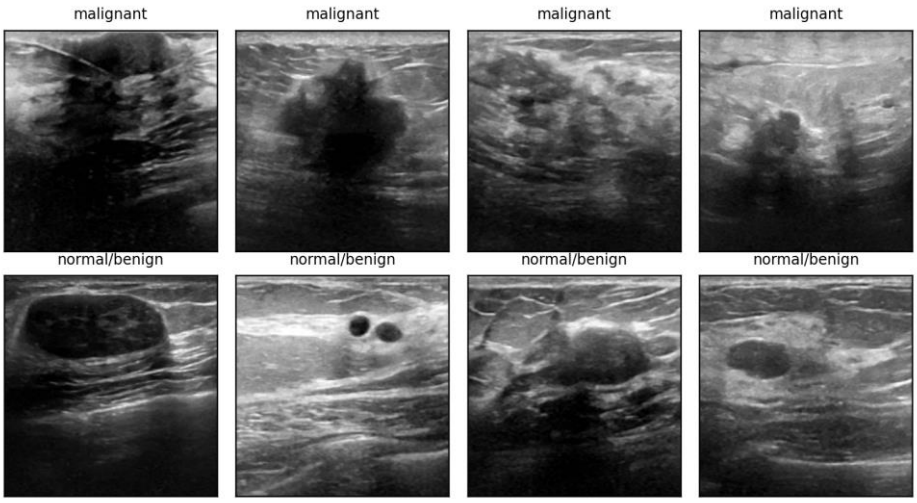

Figure 2. Representative BreastMNIST_224 test images. The examples illustrate the image modality used for classification and visual explanation

To keep the experiment reproducible and computationally lightweight, each 224 x 224 image was converted into a 32 x 32 average-pooled image grid for classification. This preserves the image-based task while allowing several classifiers to be evaluated under the same split. The original 224 x 224 image was retained for display in the explanation cards and for heatmap overlays.

The model comparison included a majority baseline, pixel logistic regression, linear SVM, RBF SVM, random forest, and a one-hidden-layer MLP. Table 3 lists the fixed settings. Class weighting was used for classifiers that support it because the malignant class is the minority class. The selected card model was the RBF SVM because it provided the strongest held-out AUROC among the evaluated models and produced smooth probability estimates through Platt scaling in the SVM probability output.

The official validation split was used to select the operating threshold for the selected model. The threshold was chosen to achieve at least 0.95 validation sensitivity while maximizing validation specificity; this produced a threshold of 0.254. This threshold is intentionally lower than 0.5 because the prototype is framed as diagnostic review support rather than autonomous diagnosis. The official test split was used for all final metrics in Tables 4 through 10.

Probability quality was measured with Brier score, negative log-likelihood, and ten-bin expected calibration error. Discrimination was measured with AUROC and AUPRC. Clinical error was reported through sensitivity, specificity, PPV, NPV, and confusion counts. Bootstrap confidence intervals for selected test metrics used 1,000 resamples of the official test split.

Table 3. Classifiers and fixed experimental settings

Model	Fixed settings	Experimental role

Majority baseline	Positive probability fixed to training malignant prevalence	Reference only
Pixel logistic	32 x 32 average-pooled image grid; StandardScaler; class_weight=balanced; liblinear; max_iter=5000	Transparent linear image baseline
Linear SVM	32 x 32 grid; StandardScaler; LinearSVC C=0.1; class_weight=balanced; sigmoid calibration with 3-fold CV	Margin classifier comparison
RBF SVM	32 x 32 grid; StandardScaler; C=10.0; gamma=scale; probability=True; class_weight=balanced	Selected explanation-card model
Random forest	300 trees; min_samples_leaf=3; class_weight=balanced; seed=42	Nonlinear ensemble comparison
MLP	32 x 32 grid; StandardScaler; hidden layer=(64); alpha=0.001; early stopping; seed=42	Neural image-grid comparison

The explanation generator is deterministic. For each test image, it computes malignant-risk probability, predicted class using the validation-selected threshold, confidence as max(p, 1 - p), entropy, risk tier, uncertainty band, and an occlusion-sensitivity heatmap. The Low-risk display tier was set conservatively at $p < 0.05$, the Review tier at $0.05 <= p <= 0.75$, and the High-risk tier at $p > 0.75$. The uncertainty bands were Indeterminate for confidence < 0.65, Watchful for 0.65 <= confidence < 0.85, and Certain for confidence >= 0.85.

The occlusion heatmap was produced by replacing local regions of the 32 x 32 image grid with the training-set mean intensity and measuring the drop in malignant-risk probability. Positive drops were normalized and overlaid on the original 224 x 224 ultrasound image. This produces a visual evidence tile suitable for UI prototyping while keeping the limitation explicit: the overlay is model evidence, not a clinical segmentation or tumor boundary.

The UI/UX evaluation used two computational checks. First, risk-tier and uncertainty-band analyses measured where false negatives, false positives, and lower-accuracy cases appeared in the card hierarchy. Second, a card communication audit scored four card variants against five required fields: probability, uncertainty, evidence, limitation, and next action. This is not a clinician usability study and should not be read as evidence of user trust, comprehension, or clinical workflow improvement.

RESULTS

The held-out model comparison is reported in Table 4 and visualized in Figure 3. At the default 0.5 threshold, the MLP had the highest accuracy, but the RBF SVM had the highest AUROC and AUPRC. Because the card interface depends on a probability display and risk ranking, the RBF SVM was selected for the explanation-card prototype. The selected model's probability reliability is shown alongside other models in Figure 4.

Table 4. Held-out model comparison at the default 0.5 threshold

Model	Accuracy	AUROC	AUPRC	F1 malignant	Sensitivity	Specificity	Brier	ECE
Majority baseline	0.731	0.500	0.269	0.000	0.000	1.000	0.197	0.000
Pixel logistic	0.821	0.829	0.588	0.689	0.738	0.851	0.160	0.142
Linear SVM	0.763	0.840	0.654	0.275	0.167	0.982	0.151	0.110
RBF SVM	0.808	0.867	0.728	0.595	0.524	0.912	0.125	0.068
Random forest	0.808	0.850	0.711	0.545	0.429	0.947	0.134	0.091
MLP	0.833	0.826	0.694	0.639	0.548	0.939	0.137	0.087

Note. The selected card model was the RBF SVM because it had the strongest held-out AUROC and AUPRC among the evaluated models.

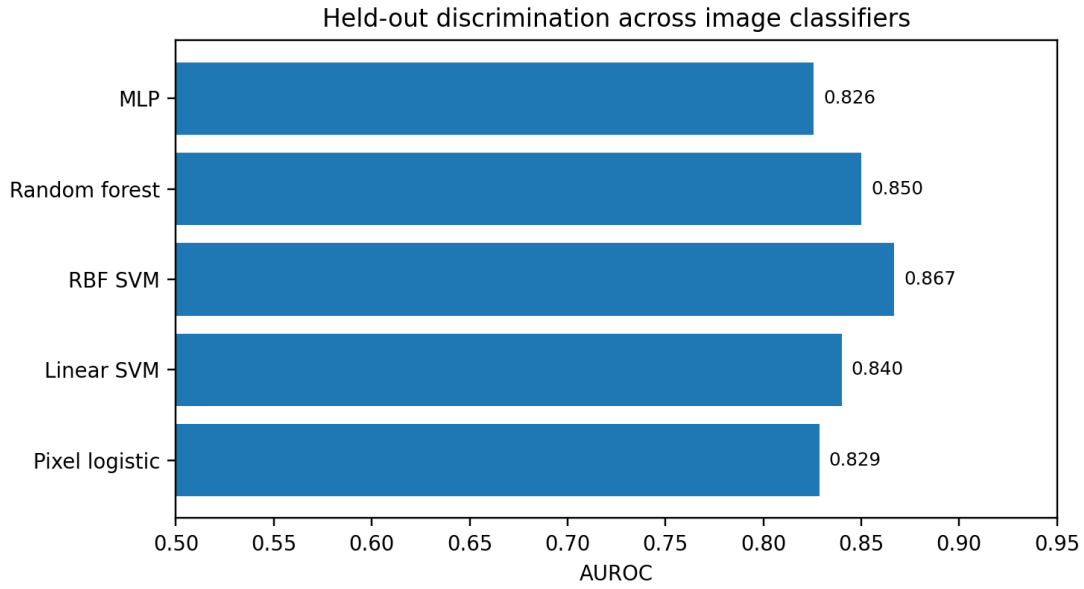

Figure 3. Held-out AUROC comparison across the non-baseline image classifiers

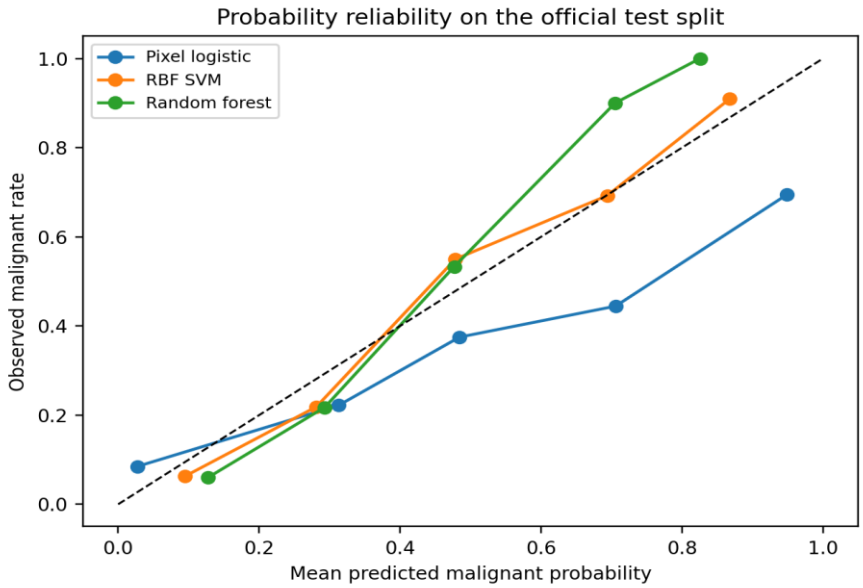

Figure 4. Reliability diagram comparing probability displays for representative image classifiers on the official test split

For the selected RBF SVM, the validation-selected operating threshold was 0.254. Table 5 reports validation and test results at this threshold, including bootstrap confidence intervals for the test split. On the official test set, the selected operating point achieved sensitivity = 0.833 and specificity = 0.746. The confusion matrix in Figure 5 and Table 6 shows that the model produced 35 true positives, 85 true negatives, 7 false negatives, and 29 false positives. The false-positive burden is expected from a sensitivity-favoring threshold and reinforces why the card should be framed as review support rather than diagnosis.

Table 5. Selected RBF SVM threshold results. The operating threshold was t = 0.254

Metric	Validation split	Test split	Test bootstrap 95% CI
Accuracy	0.885	0.769	0.705-0.833
AUROC	0.945	0.867	0.802-0.929
Sensitivity	0.952	0.833	0.702-0.941
Specificity	0.860	0.746	0.670-0.823
Brier	0.100	0.125	0.095-0.155
ECE	0.139	0.068	0.055-0.137

Table 6. Confusion counts and clinical diagnostic metrics for the selected card model on the official test split

TN	FP	FN	TP	Sensitivity	Specificity	PPV	NPV	Accuracy
85	29	7	35	0.833	0.746	0.547	0.924	0.769

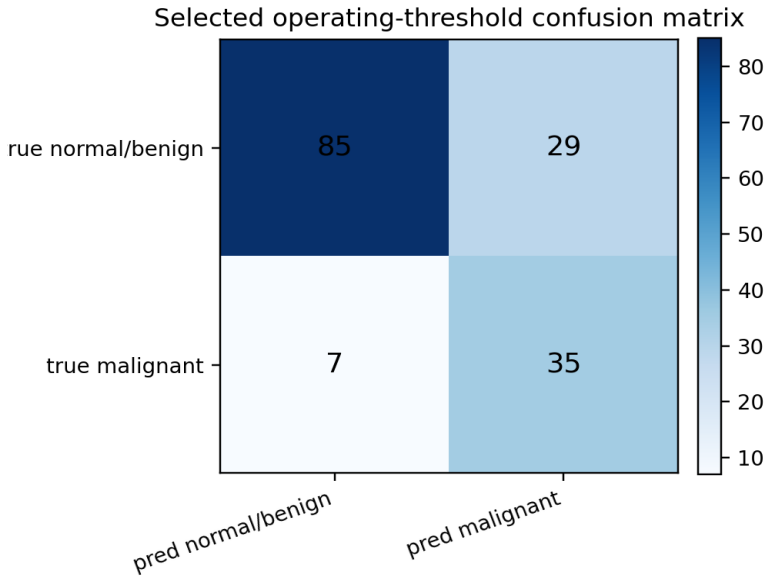

Figure 5. Confusion matrix for the selected RBF SVM card model at the validation-selected operating threshold

The card-level results show how the interface hierarchy handles uncertainty and error. Table 7 reports risk-tier behavior. The conservative Low-risk tier contained six test cases and no malignant cases. The Review tier contained most cases, including all seven false negatives. The High-risk tier contained 14 cases, 11 of which were malignant; the three high-risk false positives illustrate why the card uses risk language and follow-up cues rather than diagnostic language. Figure 6 visualizes the observed class distribution across the risk tiers.

Table 7. Explanation-card risk-tier results for the selected card model

Risk tier	N	Mean predicted risk	Observed malignant rate	Accuracy	False negatives	False positives
Low risk	6	0.042	0.000	1.000	0	0
Review	136	0.239	0.228	0.757	7	26
High risk	14	0.849	0.786	0.786	0	3

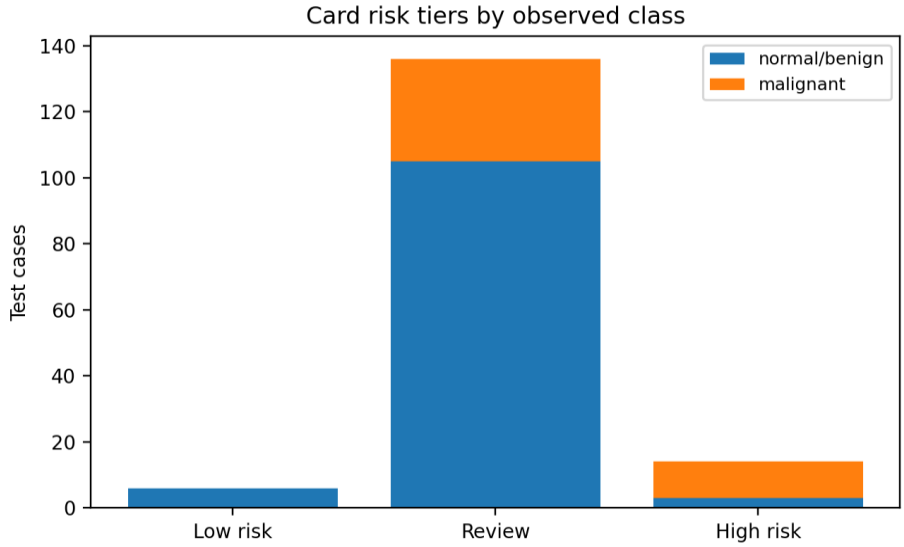

Figure 6. Card risk tiers by observed class on the official test split

Table 8 reports uncertainty-band behavior. Indeterminate and Watchful cases had lower accuracy than Certain cases, which is consistent with the goal of making ambiguity visible. The Certain band reached 0.949 accuracy, while the Indeterminate band reached 0.536 accuracy. This does not prove future reliability, but it shows that the implemented uncertainty hierarchy captured meaningful variation in this test run.

Table 8. Uncertainty-band results for the selected card model

Uncertainty band	N	Accuracy	Mean confidence	Mean entropy
Indeterminate	28	0.536	0.568	0.982
Watchful	49	0.612	0.755	0.791
Certain	79	0.949	0.912	0.423

Figure 7 shows representative occlusion-sensitivity overlays. The overlays are intentionally presented as visual model evidence rather than as anatomical annotations. The card mockup in Figure 8 demonstrates how the image evidence, probability, risk tier, uncertainty band, limitation, and next action fit into one designer-editable component. Table 9 lists representative generated cards, including true-positive, true-negative, false-negative, false-positive, and borderline cases.

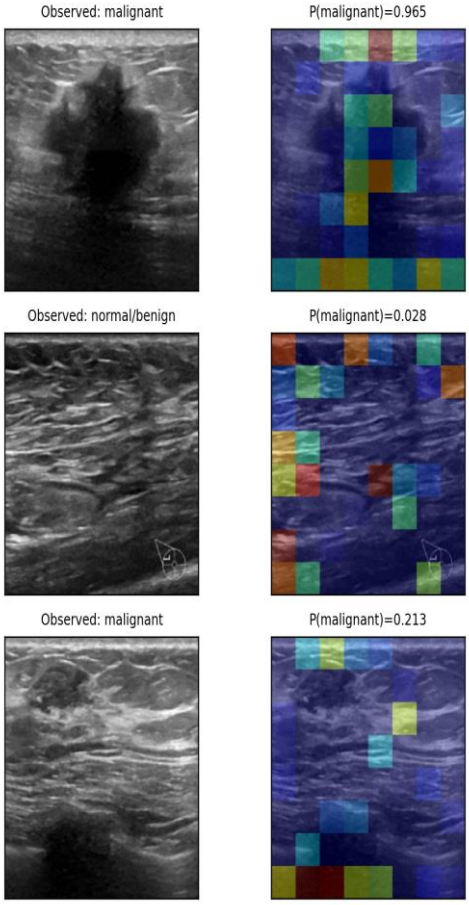

Figure 7. Representative occlusion-sensitivity heatmaps overlaid on original BreastMNIST_224 ultrasound images

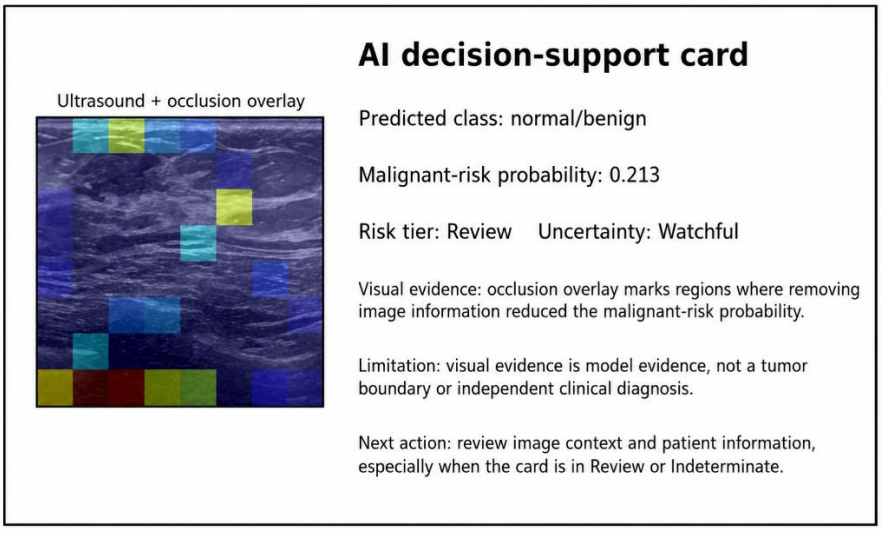

Figure 8. Explanation-card mockup showing risk hierarchy, uncertainty band, occlusion evidence, limitation statement, and review cue

Table 9. Representative generated explanation cards from the official test split

Test case	True label	Predicted label	P(malignant)	Confidence	Risk tier	Uncertainty
6	malignant	malignant	0.965	0.965	High risk	Certain
96	normal/benign	normal/benign	0.028	0.972	Low risk	Certain
68	malignant	normal/benign	0.213	0.787	Review	Watchful
75	normal/benign	malignant	0.268	0.732	Review	Watchful
1	normal/benign	normal/benign	0.253	0.747	Review	Watchful
141	malignant	normal/benign	0.094	0.906	Review	Certain

Note. Predicted labels use the validation-selected operating threshold. Risk tier and uncertainty are displayed separately to avoid treating a class label as a diagnosis.

The communication audit in Table 10 confirms why a prediction-only interface is incomplete. A card that shows only a class label and probability omits uncertainty, evidence, limitation, and next action. Adding uncertainty and safeguards improves coverage, but the full image explanation card is the first variant that contains all five required fields. These scores are content-coverage checks only and should not be interpreted as user-study outcomes.

Table 10. UI explanation-card communication audit

Card variant	Probability	Uncertainty	Evidence	Limitation	Next action	Coverage /5	Overtrust flag
A. Prediction only	1	0	0	0	0	1	1
B. Probability + uncertainty	1	1	0	1	1	4	0
C. Image explanation card	1	1	1	1	1	5	0
D. Explanation + calibration note	1	1	1	1	1	5	0

DISCUSSION

The empirical study supports a narrower and more image-appropriate claim than a broad statement about clinical diagnostic interfaces. The results show that a BreastMNIST_224 image classifier can be connected to an uncertainty-aware explanation-card structure, and that the card can display probability, risk tier, uncertainty, occlusion evidence, limitation, and follow-up

guidance in one compact visual hierarchy. The computational results do not show that clinicians will trust the card appropriately or make better decisions; they show that the interface artifact contains the information required for cautious review.

The most important design implication is that uncertainty should remain visible at the same level as risk. In the selected card model, all test-set false negatives occurred in the Review tier rather than in the Low-risk tier. This does not remove the error, but it changes the visual framing of the error. A binary prediction display would show those cases simply as normal/benign, whereas the card structure keeps them in a review-oriented zone because their risk values do not meet the conservative Low-risk cutoff.

The heatmap evidence also requires careful visual language. Occlusion heatmaps provide spatial cues about the image regions that influenced the model score, but the overlay is not a clinical localization mask. For this reason, the card text describes the overlay as model evidence and pairs it with a limitation statement. This design choice reduces the chance that a visually salient heatmap will be mistaken for a ground-truth tumor boundary.

The thresholding results highlight the difference between model performance and interface behavior. At the default 0.5 threshold, the selected RBF SVM had higher specificity but lower sensitivity. The validation-selected threshold improved sensitivity at the cost of more false positives. This trade-off is appropriate for a review-support prototype only if the interface communicates uncertainty and limits. The explanation card therefore avoids language such as 'diagnosis' and instead uses risk, review, and next-action terms.

For graphic design and UI/UX research, the practical contribution is a reusable visual grammar. The image tile and heatmap occupy the evidence zone; risk and uncertainty occupy the primary reading path; limitations and next action occupy the safety zone. The same grammar can be evaluated in future clinician-facing studies by comparing prediction-only cards, uncertainty-only cards, and full image explanation cards under controlled tasks.

Limitations

The first limitation is dataset scale. BreastMNIST_224 contains 780 images, and the official test split contains 156 images. The resulting confidence intervals are therefore wide enough that the reported metrics should be interpreted as prototype evidence rather than clinical validation. External datasets, temporal validation, and prospective workflow studies are required before any deployment claim.

The second limitation concerns the classifier. The selected model uses a downsampled 32 x 32 image grid derived from the 224 x 224 source image. This design keeps the experiment

lightweight and reproducible, but it does not exploit all image detail. A future study should compare compact CNNs, transfer learning, and segmentation-aware models while preserving the same explanation-card structure.

The third limitation is the explanation method. Occlusion sensitivity is intuitive and model-agnostic, but it is coarse and sensitive to patch size. Grad-CAM, perturbation maps, and segmentation-informed saliency methods should be compared in future work. Regardless of method, the interface should continue to label visual evidence as model evidence rather than clinical proof.

The fourth limitation is UI evaluation. The communication audit verifies information coverage, and the risk-tier analysis verifies where errors appear in the card hierarchy. Neither test measures clinician comprehension, diagnostic time, trust calibration, or reliance behavior. Claims about interface usefulness require controlled user studies with clinicians, trainees, or other intended users.

CONCLUSION

This paper presents an uncertainty-aware visual communication framework for breast ultrasound explanation cards and evaluates it on BreastMNIST_224. The selected image model achieved AUROC = 0.867 on the official test split, and its validation-selected operating threshold achieved sensitivity = 0.833 and specificity = 0.746. The card prototype adds risk tier, uncertainty band, occlusion heatmap evidence, limitation text, and next action to the model output. In the test split, the conservative Low-risk tier contained no malignant cases, and all false negatives remained in the Review tier. These results support a cautious prototype-level conclusion: diagnostic AI interfaces should present risk, uncertainty, visual evidence, limits, and follow-up guidance together, and they should avoid presenting a class label as a standalone diagnosis.

REFERENCES

Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2020). Dataset of breast ultrasound images. Data in Brief, 28, 104863. https://doi.org/10.1016/j.dib.2019.104863

Amershi, S., Weld, D., Vorst, M., Chilton, L., Kim, J., Ruamviboonsuk, P., & Horvitz, E. (2019). Guidelines for human-AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-13. https://doi.org/10.1145/3290605.3300233

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). Hello AI: Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making.

Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-24. https://doi.org/10.1145/3359206

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://arxiv.org/abs/1702.08608

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI - Explainable artificial intelligence. Science Robotics, 4(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

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.

Jason Kuhn, Yushan Chen, & Evelyn Chan. (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

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17, 195. https://doi.org/10.1186/s12916-019-1426-2

Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31-57. https://doi.org/10.1145/3236386.3241340

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

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

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. Proceedings of the 22nd International Conference on Machine Learning, 625-632. https://doi.org/10.1145/1102351.1102430

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

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

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. International Conference on Learning Representations Workshop.

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. Proceedings of the Machine Learning for Healthcare Conference, 106-124.

Yang, J., Shi, R., & Ni, B. (2021). MedMNIST classification decathlon: A lightweight AutoML benchmark for medical image analysis. 2021 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, 41. https://doi.org/10.1038/s41597-022-01721-8

Yushan Chen, & Evelyn Chan. (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

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. European Conference on Computer Vision, 818-833. https://doi.org/10.1007/978-3-319-10590-1_53