PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 04 (2026) Page: 48-57
E-ISSN: 3109-8886

Uncertainty-Aware Lightweight Breast Ultrasound Modeling on BreastMNIST+: Parameter-Efficient Label-Prompt Alignment, Controlled-Shift Robustness, and Verifiable Evidence Cards

Jiaqi Sun[1], Jiaqi Sun[2], Tyler Henderson[3], Kevin Zhao[4], Ashley Cooper[5]

[1]Department of Computer Science, The University of Texas at Austin, Austin, TX, USA
[2]Department of Computer Science, University of Washington, Seattle, WA, USA
[3]Department of Computer Science, University of California San Diego, La Jolla, CA, USA
[4]Department of Computer Science, Purdue University, West Lafayette, IN, USA

Email : [1*]jq.sun88@outlook.com

Abstract

On BreastMNIST+ 224, we tested visual and class-prompt models with calibration, conformal sets, routing, perturbations, and evidence cards. The 546/78/156 train/validation/test artifact contained one conflicting duplicate. The ensemble achieved AUROC/AUPRC 0.885/0.738, balanced accuracy 0.842, sensitivity 0.833, and specificity 0.851. Temperature scaling reduced Brier/ECE 0.120/0.093-to-0.116/0.079. Label-conditional 90% conformal prediction achieved overall/malignant coverage 0.897/0.952 and size 1.359. Rank-2 prompts reached AUROC 0.790; blending raised sensitivity to 0.905 but lowered specificity to 0.763. Rotation and contrast degraded most. 156 cards passed checks. Results support routing, not clinical validity.

Keywords: Breast Ultrasound, Medmnist+, Medical Artificial Intelligence, Label-Prompt Alignment, Reduced-Rank Modeling, Probability Calibration, Conformal Prediction, Selective Classification, Distribution Shift, Evidence Cards.

Abstrak

Pada BreastMNIST+ 224, kami menguji model visual dan class-prompt dengan menerapkan kalibrasi, conformal sets, routing, perturbasi, dan evidence cards. Kumpulan data latih/validasi/uji yang terdiri dari 546/78/156 sampel tersebut memuat satu duplikat yang saling bertentangan. Model ensemble mencapai nilai AUROC/AUPRC sebesar 0,885/0,738, akurasi seimbang 0,842, sensitivitas 0,833, dan spesifisitas 0,851. Temperature scaling menurunkan nilai Brier/ECE dari 0,120/0,093 menjadi 0,116/0,079. Metode conformal prediction 90% yang bergantung pada label (label-conditional) mencapai cakupan keseluruhan/ganas sebesar 0,897/0,952 dengan ukuran set 1,359. Prompt peringkat-2 (rank-2) mencapai AUROC 0,790; teknik blending meningkatkan sensitivitas menjadi 0,905 namun menurunkan spesifisitas menjadi 0,763. Rotasi dan kontras menyebabkan penurunan performa paling signifikan. Sebanyak 156 evidence cards lolos pemeriksaan. Hasil penelitian mendukung penggunaan mekanisme routing, namun tidak secara langsung mendukung validitas klinis.

***Kata Kunci**: USG Payudara, Medmnist+, Kecerdasan Buatan Medis, Penyelarasan Label-Prompt, Pemodelan Peringkat Tereduksi, Kalibrasi Probabilitas, Conformal Prediction, Klasifikasi Selektif, Pergeseran Distribusi, Evidence Cards.*

A. INTRODUCTION

Small medical-image benchmarks magnify overfitting, calibration instability, and data errors; split audits, complementary metrics, intervals, and bounded claims are therefore essential [1], [2].

MedMNIST fixes biomedical splits [3], while MedMNIST+ adds multiple resolutions [4]. BreastMNIST derives from BUSI ultrasound [5], merging normal/benign against malignant. Resolution [6] and corruption [7] studies contextualize an image-label artifact without reports, identifiers, demographics, sites, outcomes, or masks.

MobileNetV2 [8] and patch transformers [9] motivate compact branches. MedMNIST transfer [10] and calibrated vision-language work [11] are relevant, but label-only supervision is not image-text pretraining.

Contrastive [12], self-distilled [13], and biomedical transfer [14] motivate invariance; LoRA [15] and adapters [16] motivate economical mappings, not foundation-model adaptation here.

CLIP [17], MedCLIP [18], GLoRIA [19], biomedical foundation models [20], and literature-grounded models [21] use real language supervision; fixed BreastMNIST phrases do not. DenseNet pathology [22] and MRD prediction [23] are distinct oncology settings.

Temperature scaling [24], Brier scoring [25], calibrated fusion [26], and ensembles [27] address probability quality.

Diterima Redaksi: 01-08-2026 | Selesai Revisi: 06-08-2026 | Diterbitkan Online: 07-08-2026

Uncertainty can fail under shift [28] or broken shared features [29].

Conformal sets [30], [31] and selective classification [32] support uncertainty-aware routing. Cross-domain calibration, explanation [33], rejection [34], and imbalance evaluation [35] are workflow analogues, not clinical evidence.

Corruptions do not establish site robustness [36], and saliency requires caution [37]. Model cards [38], medical interfaces [39], BreastMNIST cards [40], and patient-facing frameworks [41], [42] motivate constrained outputs.

Evidence cards [43], provenance RAG [44], and calibrated abstention [45] motivate inspectability. Here fields are reconstructed from arrays; no clinical claim is generated.

We compare supervised, unlabeled-view, and reduced-rank models, then link calibration, Mondrian sets, routing, perturbations, and validated cards. The aim is benchmark evaluation, not clinical assessment.

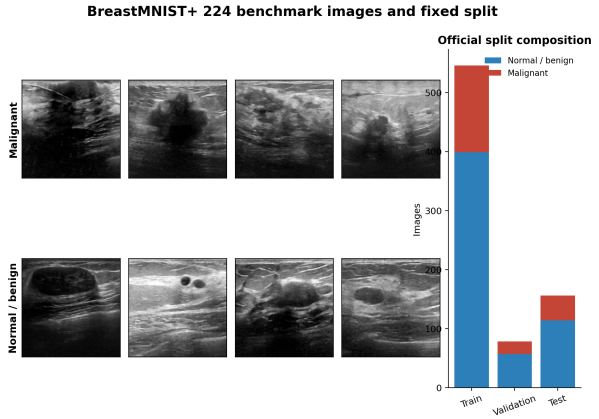

Figure 1 Representative Malignant And Normal-Or-Benign Breastmnist+ Test Images With Fixed Split Counts

B. METHODS

Data, Label Convention, And Integrity Audit

We used the BreastMNIST+ 224-pixel v3.0 NPZ [4]; MD5 b56378a6eefa9fed602bb16d192d4c8b matched metadata. Six arrays held grayscale images and labels. Intensities were divided by 255; because malignant is official label 0 [3], y_malignant = 1 - official_label.

Train/validation/test contained 546/78/156 images, each 26.923% malignant. SHA-256 checks found no within-split duplicates or train-validation/validation-test overlap. One train-test image conflicted (train 115 malignant; test 76 normal/benign). We retained the official split, then excluded test 76 and, separately, both members with refitting.

Table 1 Dataset Composition And Integrity-Audited Splits

Split	N	Malignant	Normal / benign	Malignant %	Pixel mean	Pixel SD
Training	546	147	399	26.9	0.328	0.219
Validation	78	21	57	26.9	0.339	0.223
Test	156	42	114	26.9	0.336	0.218
Total	780	210	570	26.9	0.330	0.219

Visual Representations And Supervised Baselines

Pixel-LR used 28-by-28 pooled pixels and balanced L2 logistic regression (C=0.1). PCA64-RBF used 64 whitened components and balanced RBF-SVM (C=1). HOG56-RBF used 56-by-56 images, nine bins in 7-by-7 cells, 576 standardized features, and the same SVM.

Patch-MLP combined regional intensity, gradient, and Laplacian summaries with pooled pixels, retained 64 whitened components, and fitted a 16-unit ReLU MLP with Adam, balanced weights, L2=0.02, batch 64, early stopping, and 800-iteration maximum.

Log14-ExtraTrees used log-transformed 14-by-14 intensities and three balanced 400-tree ensembles. Visual-Ensemble averaged PCA64-RBF and tree probabilities. Runtime included features; latency averaged three raw-image passes, and size included inference objects.

Unlabeled-View Representation Learning And Class-Prompt Alignment

The label-free encoder formed two augmented 28-by-28 views. Whitened PCA and cross-covariance SVD produced a 48-dimensional average projection for balanced RBF-SVM classification, following contrastive invariance [12] and calibration-light transfer [14].

Three phrase variants per class became 64-dimensional character-ngram TF-IDF prototypes without a pretrained language model. Reduced-rank regression mapped image PCA features to prototypes; SVD tested ranks 1/2/4/8 and cosine scores used a scale-5 sigmoid. Two prototypes had no semantic rank beyond 2. This applies LoRA/adapter ideas [15], [16] at MedMNIST scale [10], not vision-language adaptation. Blend weights were 0.45/0.45/0.10 for PCA64-RBF, trees, and rank-2 prompts.

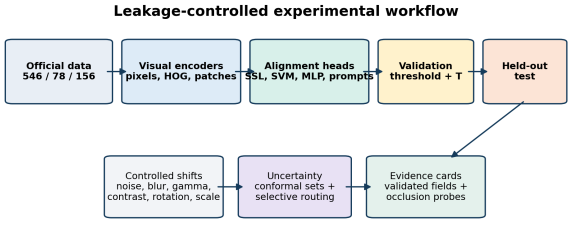

Figure 2 Leakage-Controlled Workflow From The Official Split Through Encoders, Alignment Heads, Uncertainty Analysis, Shifts, And Evidence Cards.

Table 2 Model Configuration, End-To-End Runtime, And Serialized Footprint

Model	Implementation	Fit s	ms / case	Size KB
Majority	Training prevalence	0.00	0.000	0.0
Pixel-LR	28x28 pixels + balanced LR	0.10	0.429	25.4
PCA64-RBF	PCA-64 + balanced RBF-SVM	0.32	0.483	913.3
HOG56-RBF	56x56 HOG + RBF-SVM	0.41	1.346	2279.2
Patch-MLP	Patch statistics + PCA + MLP	3.86	8.088	554.9
SSL-View48-RBF	Two-view 48-D SSL + RBF-SVM	0.66	0.784	756.4
PromptAlign-r1	Rank-1 image-to-label alignment	0.34	0.458	223.8
PromptAlign-r2	Rank-2 image-to-label alignment	0.34	0.483	224.8
PromptAlign-r4	Rank-4 image-to-label alignment	0.34	0.618	226.8
PromptAlign-r8	Rank-8 image-to-label alignment	0.34	0.710	230.8
Log14-ExtraTrees	14x14 log pixels + Extra Trees	11.32	2.833	14847.4
Visual-Ensemble	Equal visual blend	11.65	2.762	15760.7
LabelPrompt-Blend	Visual blend + 10% prompt branch	11.99	4.480	15985.6

Training, Thresholds, Metrics, And Statistical Analysis

Seeds 13/29/47 were averaged. Test data never informed fitting, calibration, prompts, or thresholds. Validation balanced accuracy set thresholds, with malignant sensitivity breaking ties. The majority baseline predicted every case normal/benign.

Primary metrics were AUROC and malignant AUPRC; 42/156 test cases were malignant, making precision-recall essential [35], [46]. We also report threshold metrics, confusion counts, NLL, Brier [25], and descriptive ten-bin ECE.

Compact IoT inference [47], edge/cloud routing [48], cold-start adaptation [49], wearable transfer [50], cross-cloud transfer [51], and regime-sensitive forecasting [52] motivate footprint and fixed-policy shift reporting, not medical comparison.

Across three seeds, 2,000 stratified paired bootstraps produced percentile 95% intervals and Holm-adjusted AUROC contrasts; exact McNemar tests compared errors.

Calibration, Conformal Prediction, And Selective Routing

A positive temperature minimized validation NLL and was applied unchanged to test logits [24]. Platt and isotonic were comparators; validation reset thresholds. This separates prediction, calibration, and decisions [26], [33].

Nonconformity was 1 - p(y_true|x); alpha 0.10/0.20 quantiles used ceil((n + 1)(1 - alpha)). Marginal sets used one quantile; Mondrian sets used classwise quantiles from 21 malignant and 57 normal/benign cases. We report coverage, size, and set-type rates. This follows conformal classification [30], [31] and workflow separation [33]; shifted coverage is measured, not guaranteed.

Operational conformal envelopes [53], alert control [54], evidence attribution [55], and counterfactual workflows [56] separate calibrated risk from explanation; all are nonmedical analogues.

Validation-fixed decision-margin cutoffs targeted 90/80/70/60% retention. We measured achieved coverage, error, balanced accuracy, malignant sensitivity, and risk-coverage area. This exposes abstention cost [32]; credit-risk rejection [34] is only an analogue.

Controlled Perturbations And Evidence-Card Validation

Fixed-seed test copies received noise, blur, gamma, contrast, rotations, or downsample/restore. Clean calibration, thresholds, and conformal quantiles remained fixed while all metrics were recomputed. These are corruption probes [7], [36], not scanner or hospital validation.

Cards stored labels, prediction, calibrated score, 90% Mondrian set, routing, perturbation range, occlusion regions, version, and warning. Occlusion measured score sensitivity, not lesions [37]. The schema combined model cards [38], medical interfaces [39]-[42], evidence cards [43], provenance RAG [44], [45], and deterministic narratives [57]. Every field was recomputed and audited.

C. RESULTS AND DISCUSSION

Clean-Test Discrimination And Reduced-Rank Alignment

The majority rule had zero malignant sensitivity (Table III). Pixel-LR, PCA64-RBF, and trees reached AUROC/AUPRC 0.822/0.573, 0.865/0.690, and 0.867/0.725. Their ensemble led at 0.885/0.738, balanced accuracy 0.842, sensitivity 0.833, specificity 0.851, and TN/FP/FN/TP=97/17/7/35.

Patch-MLP reached AUROC/AUPRC 0.855/0.729; SSL reached AUROC 0.850. This limits inference to the tested objective, not DINO [13]. Transfer is domain-dependent [14], and customer representations [58] are only an analogue. Ensemble seed AUROCs were 0.883-0.887.

Table 3 Clean Official-Test Performance With Malignancy As The Positive Class

Model	AUROC	AUPRC	Bal. acc.	Sens.	Spec.	Brier	ECE
Majority	0.500	0.269	0.500	0.000	1.000	0.197	0.094
Pixel-LR	0.822	0.573	0.743	0.714	0.772	0.159	0.119
PCA64-RBF	0.865	0.690	0.810	0.786	0.833	0.127	0.058
HOG56-RBF	0.692	0.575	0.682	0.548	0.816	0.169	0.106
Patch-MLP	0.855	0.729	0.769	0.714	0.825	0.133	0.087
SSL-View48-RBF	0.850	0.681	0.788	0.786	0.789	0.131	0.086
PromptAlign-r2	0.790	0.581	0.679	0.524	0.833	0.266	0.243
Log14-ExtraTrees	0.867	0.725	0.815	0.762	0.868	0.124	0.081
Visual-Ensemble	0.885	0.738	0.842	0.833	0.851	0.120	0.093
LabelPrompt-Blend	0.879	0.722	0.834	0.905	0.763	0.123	0.093

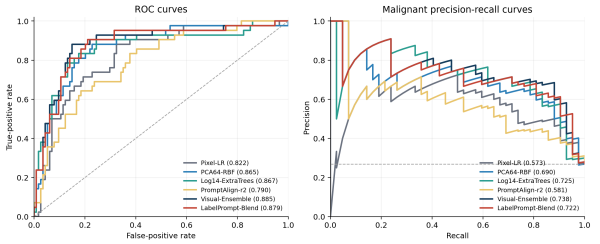

Figure 3 Official-Test ROC And Malignant-Class Precision-Recall Curves. Legend Values Are AUROC And AUPRC, Respectively.

Prompt rank 1 reached AUROC 0.658; ranks 2/4/8 tied at 0.790. The blend reached AUROC/AUPRC 0.879/0.722, raising sensitivity 0.833-to-0.905 while lowering specificity 0.851-to-0.763. It changed operating balance, not discrimination, and is not image-text pretraining [18]-[21] or MedMNIST transfer [10], [11].

Table 4 Temperature-Scaled Primary Ensemble Bootstrap Intervals And Paired Auroc Contrasts

Analysis	Estimate	95% interval	Adjusted p
Primary AUROC	0.885	0.817-0.941	-
Primary AUPRC	0.738	0.620-0.864	-
Temperature-scaled Brier	0.116	0.088-0.147	-
Primary Balanced accuracy	0.842	0.775-0.900	-
Primary Sensitivity	0.833	0.714-0.929	-
Primary Specificity	0.851	0.781-0.912	-
AUROC vs PCA64-RBF	0.020	-0.012-0.056	0.420
AUROC vs Log14-ExtraTrees	0.019	-0.011-0.048	0.420

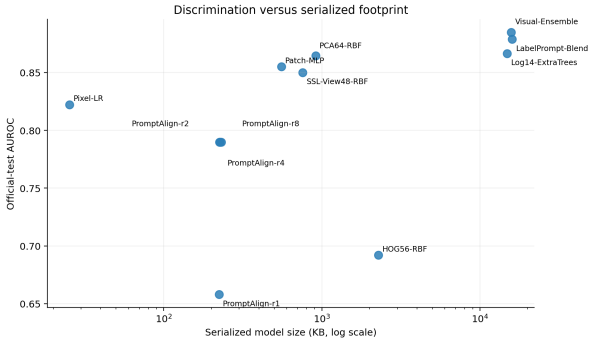

Figure 4 Official-Test AUROC Versus Serialized Standalone Model Footprint.

Ensemble AUROC/AUPRC intervals were 0.817-0.941/0.620-0.864 (Table IV). AUROC gains over PCA64-RBF and trees were 0.020 and 0.019, with intervals crossing zero and adjusted p=0.420. McNemar tests found no error difference; point estimates did not establish superiority.

Integrity sensitivity and calibration

Excluding test 76 raised AUROC/AUPRC 0.885/0.738-to-0.892/0.784. Removing both members and refitting yielded AUROC 0.892, AUPRC 0.790, specificity 0.876, and Brier 0.110 on 155 cases (Table V). Conclusions were unchanged.

Table 5 Temperature-Scaled Ensemble Sensitivity To The Conflicting Train-Test Duplicate

Evaluation	AUROC	AUPRC	Bal. acc.	Sens.	Spec.	Brier	ECE
Official test, n=156	0.885	0.738	0.842	0.833	0.851	0.116	0.079
Exclude test conflict, n=155	0.892	0.784	0.846	0.833	0.858	0.111	0.067
Refit and exclude conflict, 545/155	0.892	0.790	0.843	0.810	0.876	0.110	0.054

Temperature T=0.662 reduced Brier/NLL/ECE 0.120/0.390/0.093-to-0.116/0.377/0.079 without changing ranks or decisions (Table VI). Platt reached Brier 0.115 and ECE 0.073. Isotonic ties lowered AUROC to 0.873 and raised NLL to 0.808.

Table 6 Calibration Of The Primary Visual Ensemble

Calibration	T	Brier	NLL	ECE	Sens.	Spec.
Raw	-	0.120	0.390	0.093	0.833	0.851
Temperature	0.662	0.116	0.377	0.079	0.833	0.851
Platt	-	0.115	0.377	0.073	0.833	0.851
Isotonic	-	0.126	0.808	0.102	0.833	0.851

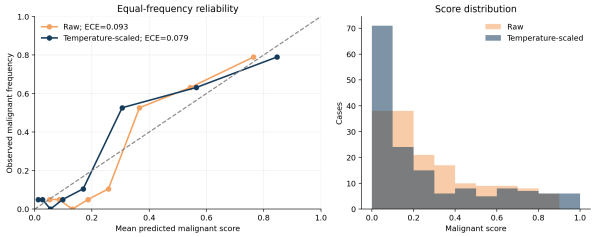

Figure 5 Equal-Frequency Reliability And Score Distributions Before And After Temperature Scaling.

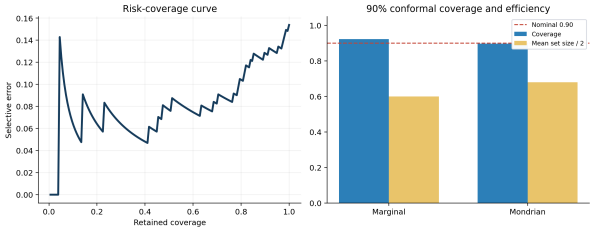

Figure 6 Risk-Coverage Curve And 90% Conformal Coverage-Set-Size Trade-Off.

Prediction Sets And Selective Routing

At nominal 90%, marginal coverage/size was 0.923/1.199, but malignant coverage was 0.810. Mondrian overall/malignant/normal coverage was 0.897/0.952/0.877 with size 1.359. Fifty-six cases received both labels and none an empty set (Table VII).

At nominal 80%, marginal overall/malignant coverage fell to 0.737/0.357 with 27 empty sets. Mondrian coverage was 0.821 overall and 0.881 malignant, with no empty sets and size 1.013. Marginal summaries concealed minority failure; guarantees are not individualized.

Table 7 Marginal And Mondrian Conformal Prediction Sets

Method	Nominal	Overall	Mal. cov.	N/B cov.	Mean size	Singleton	Two-label	Empty
Marginal	0.900	0.923	0.810	0.965	1.199	0.801	0.199	0.000
Mondrian	0.900	0.897	0.952	0.877	1.359	0.641	0.359	0.000
Marginal	0.800	0.737	0.357	0.877	0.827	0.827	0.000	0.173
Mondrian	0.800	0.821	0.881	0.798	1.013	0.987	0.013	0.000

Routing reduced error throughout (Table VIII). At the 80% target, 122/156 cases remained, error fell 0.154-to-0.090, balanced accuracy was 0.891, and malignant sensitivity 0.857. At 70%, 104 cases remained with error 0.077 and sensitivity 0.870. AURC was 0.082.

Table 8 Selective Routing With Validation-Fixed Decision-Margin Cutoffs

Val. target	Test cov.	N	Error	Bal. acc.	Mal. sens.	Margin cutoff	AURC
1.000	1.000	156	0.154	0.842	0.833	0.000	0.082
0.900	0.891	139	0.122	0.871	0.857	0.143	0.082
0.800	0.782	122	0.090	0.891	0.857	0.365	0.082
0.700	0.667	104	0.077	0.904	0.870	0.510	0.082
0.600	0.577	90	0.078	0.889	0.833	0.600	0.082

Controlled-Shift Robustness

Noise and blur changed little (Table IX): noise SD 0.06 yielded AUROC/balanced accuracy 0.883/0.846; blur sigma 2 yielded 0.883/0.853. Downsampling preserved rounded metrics because branches averaged no finer than 28-by-28; arbitrary information loss is not harmless.

Gamma 1.50 reduced AUROC/balanced accuracy to 0.847/0.773; contrast 0.50 yielded AUROC 0.854, Brier 0.136, and coverage 0.788. Fifteen-degree rotation was worst: AUROC 0.821, Brier 0.141, size 1.532, and coverage 0.865. Calibration can fail under shift [28], [29].

Table 9 Controlled-Perturbation Performance And Uncertainty

Condition	AUROC	AUPRC	Bal. acc.	Brier	ECE	90% cov.	Set size
Clean	0.885	0.738	0.842	0.116	0.079	0.897	1.359
Gaussian noise, sigma=0.03	0.885	0.738	0.842	0.116	0.078	0.891	1.353
Gaussian noise, sigma=0.06	0.883	0.746	0.846	0.118	0.077	0.891	1.353
Gaussian blur, sigma=1.0	0.884	0.736	0.842	0.116	0.078	0.897	1.359
Gaussian blur, sigma=2.0	0.883	0.739	0.853	0.116	0.076	0.891	1.353
Gamma, gamma=0.75	0.879	0.754	0.806	0.122	0.082	0.904	1.449
Gamma, gamma=1.50	0.847	0.705	0.773	0.133	0.074	0.853	1.462
Contrast, factor=0.75	0.881	0.742	0.833	0.117	0.080	0.853	1.391
Contrast, factor=0.50	0.854	0.680	0.801	0.136	0.083	0.788	1.481
Rotation, +/-7 deg	0.857	0.717	0.759	0.131	0.062	0.878	1.462
Rotation, +/-15 deg	0.821	0.686	0.778	0.141	0.058	0.865	1.532
Resolution, 112 px	0.885	0.738	0.842	0.116	0.079	0.897	1.359
Resolution, 56 px	0.885	0.738	0.842	0.116	0.079	0.897	1.359

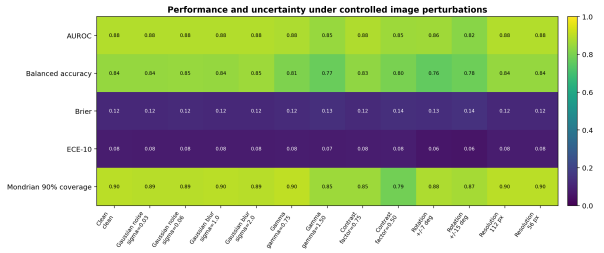

Figure 7 Measured performance and uncertainty under controlled perturbations; lower Brier and ECE values are preferable.

Evidence Cards, Limitations, And Research Implications

All 156 cards passed (Table X): 85 accept, 71 review, 56 two-label. Every field matched source arrays. The schema is narrower than radiology and BreastMNIST card work [39], [40] and follows risk-language constraints [41], [42]. Occlusion overlays are not lesions.

Table 10 Evidence-Card Consistency And Safety Audit

Audit item	Measured result
Cards generated	156
Schema and source-consistency checks passed	156
Validation pass rate	1.000
Accept cards	85
Review cards	71
Two-label conformal cards	56
Unsupported-term violations	0

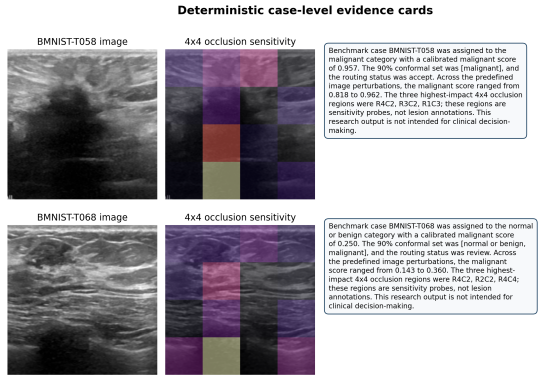

Figure 8 Deterministic Evidence-Card Examples With Coarse Occlusion-Sensitivity Probes; Overlays Are Not Lesion Annotations.

The audit tested provenance, not comprehension. Evidence cards [43]-[45], oversight cards [59], DevOps RAG [60], [61], budgeted retrieval [62], self-verification [63], and claim checking [64] separate evidence, generation, and validation. These are nonmedical precedents; schema validity does not establish understanding.

Cross-domain analogues pair calibration with refusal [65], prioritize minority-event review [66], predict failure before release [67], and abstain under transfer [68].

Deployment analogues report footprint [69], route difficult cases to costly processing [70], test evidence faithfulness [71], and expose refusal-utility trade-offs [72].

Deterministic controls check claims [73], combine grounding with approval gates [74], preserve private interaction data [75], and keep AI subordinate to qualified care [76].

Reader-study hypotheses include evidence/action cards [77], counseling handoff [78], risk-detail hierarchy [79], and human comparison of AI critique [80]. Automated validation cannot establish comprehension.

Limitations include 156 test images, 42 malignant cases, no identifiers, merged normal/benign labels, and no site, device, demographic, or temporal variation. Prompts are not reports; ECE is bin-dependent; coverage is not individualized; one validation split supports several analyses; and cards cannot confer safety. Guidance [1], [2] and distinct oncology tasks [22], [23] bound interpretation.

D. CONCLUSIONS

The ensemble achieved AUROC 0.885 and balanced accuracy 0.842; temperature scaling improved probability quality. Mondrian sets raised malignant coverage with larger sets, and routing reduced error. Prompt alignment underperformed; blending traded specificity for sensitivity. Rotation and contrast degraded performance. The duplicate and card audit show benchmark trust requires data and output controls, not clinical use.

E. REFERENCES

[1] J. Mongan, L. Moy, and C. E. Kahn, Jr., "Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers," Radiology: Artificial Intelligence, vol. 2, no. 2, art. e200029, 2020, doi: 10.1148/ryai.2020200029.

[2] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: Methodological failures and recommendations for the future," npj Digital Medicine, vol. 5, art. 48, 2022, doi: 10.1038/s41746-022-00592-y.

[3] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification," Scientific Data, vol. 10, art. 41, 2023, doi: 10.1038/s41597-022-01721-8.

[4] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "[MedMNIST+] 18x standardized datasets for 2D and 3D biomedical image classification with multiple size options: 28 (MNIST-Like), 64, 128, and 224," Zenodo, ver. 3.0, Jan. 2024, doi: 10.5281/zenodo.10519652.

[5] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," Data in Brief, vol. 28, art. 104863, 2020, doi: 10.1016/j.dib.2019.104863.

[6] S. Doerrich, F. Di Salvo, J. Brockmann, and C. Ledig, "Rethinking model prototyping through the

MedMNIST+ dataset collection," Scientific Reports, vol. 15, art. 7669, 2025, doi: 10.1038/s41598-025-92156-9.

[7] F. Di Salvo, S. Doerrich, and C. Ledig, "MedMNIST-C: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions," arXiv:2406.17536, 2024, doi: 10.48550/arXiv.2406.17536.

[8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2018, pp. 4510-4520, doi: 10.1109/CVPR.2018.00474.

A. Dosovitskiy et al., "An image is worth 16 x 16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.

[9] G. Mi, T. Ye, and D. Wood, "A lightweight medical foundation model for cross-modal multi-task pretraining and parameter-efficient few-shot transfer on MedMNIST," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 572-589, 2025, doi: 10.51903/jtie.v4i3.492.

[10] S. Lu and T. Zou, "Uncertainty-aware medical vision-language classification on a lightweight MedMNIST-compatible biomedical patch benchmark," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 1-19, 2026, doi: 10.51903/jtie.v5i2.530.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in Proc. 37th Int. Conf. Machine Learning, vol. 119, 2020, pp. 1597-1607.

[12] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in Proc. IEEE/CVF Int. Conf. Computer Vision, 2021, pp. 9650-9660, doi: 10.1109/ICCV48922.2021.00951.

[13] Q. Wu, G. Mi, and D. Wood, "Calibration-light subject-independent motor imagery BCI via self-supervised pretraining and Conformer," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 239-262, 2025, doi: 10.51903/jtie.v5i1.493.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in Proc. ICLR, 2022.

[15] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in Proc. 36th Int. Conf. Machine Learning, vol. 97, 2019, pp. 2790-2799.

A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. 38th Int. Conf. Machine Learning, vol. 139, 2021, pp. 8748-8763.

[16] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," in Proc. EMNLP, 2022, pp. 3876-3887, doi: 10.18653/v1/2022.emnlp-main.256.

[17] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition," in Proc. IEEE/CVF Int. Conf. Computer Vision, 2021, pp. 3942-3951, doi: 10.1109/ICCV48922.2021.00391.

[18] K. Zhang et al., "A generalist vision-language foundation model for diverse biomedical tasks," Nature Medicine, vol. 30, no. 11, pp. 3129-3141, 2024, doi: 10.1038/s41591-024-03185-2.

[19] C. Kim, S. U. Gadgil, A. J. DeGrave, J. A. Omiye, Z. R. Cai, R. Daneshjou, and S.-I. Lee, "Transparent medical image AI via an image-text foundation model grounded in medical literature," Nature Medicine, vol. 30, pp. 1154-1165, 2024, doi: 10.1038/s41591-024-02887-x.

[20] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, "Cancer image classification based on DenseNet model," J. Phys.: Conf. Ser., vol. 1651, no. 1, art. 012143, 2020, doi: 10.1088/1742-6596/1651/1/012143.

[21] J. Chen, J. Xiong, Y. Wang, Q. Xin, and H. Zhou, "Implementation of an AI-based MRD evaluation and prediction model for multiple myeloma," Front. Comput. Intell. Syst., vol. 6, no. 3, pp. 127-131, 2024, doi: 10.54097/zJ4MnbWW.

[22] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Machine Learning, vol. 70, 2017, pp. 1321-1330.

[23] G. W. Brier, "Verification of forecasts expressed in terms of probability," Monthly Weather Review, vol. 78, no. 1, pp. 1-3, 1950, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[24] Q. Xin, "Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning)," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 215-238, 2025, doi: 10.51903/jtie.v4i1.485.

[25] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 6402-6413.

[26] Y. Ovadia et al., "Can you trust your model's uncertainty? Evaluating predictive uncertainty

under dataset shift," in Advances in Neural Information Processing Systems, vol. 32, 2019, pp. 13991-14002.

[27] Z. S. Zhong, X. Pan, and Q. Lei, "Bridging domains with approximately shared features," in Proc. 28th Int. Conf. Artificial Intelligence and Statistics, PMLR, vol. 258, pp. 559-567, 2025.

[28] Y. Romano, M. Sesia, and E. J. Candès, "Classification with valid and adaptive coverage," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 3581-3591.

A. N. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan, "Uncertainty sets for image classifiers using conformal prediction," in Proc. ICLR, 2021.

[29] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4878-4887.

[30] J. Jin, T. Huang, and S. Lu, "A model-risk-friendly probability of default workflow: Calibration, distribution-free uncertainty quantification, and SHAP explanations on the UCI credit card default dataset," J. Adv. Comput. Syst., vol. 4, no. 6, pp. 74-85, 2024, doi: 10.69987/JACS.2024.40606.

[31] Y. Chen, Y. Zhang, D. Chau, and M. Sherman, "Credit card default risk tiering with probability calibration and uncertainty-driven rejection: A reproducible study on the UCI credit card clients dataset," J. Adv. Comput. Syst., vol. 3, no. 4, pp. 31-47, 2023, doi: 10.69987/JACS.2023.30403.

[32] J. Jin, T. Huang, and S. Lu, "Cost-sensitive learning, simulated PU learning, and one-class autoencoding for extreme-imbalance credit card fraud detection," J. Adv. Comput. Syst., vol. 4, no. 6, pp. 64-73, 2024, doi: 10.69987/JACS.2024.40605.

[33] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in Proc. ICLR, 2019.

[34] N. Arun et al., "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging," Radiology: Artificial Intelligence, vol. 3, no. 6, art. e200267, 2021, doi: 10.1148/ryai.2021200267.

[35] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in Proc. Conf. Fairness, Accountability, and Transparency, 2019, pp. 220-229, doi: 10.1145/3287560.3287596.

[36] Z. S. Zhong, Q. Wu, and G. Mi, "Uncertainty-aware medical image explanation cards: LLM-generated visual explanations for AI-assisted radiology interfaces," Int. J. Graph. Des., vol. 3, no. 2, pp. 415-436, 2025, doi: 10.51903/ijgd.v3i2.3616.

[37] T. Ye, X. Chang, and E. Zhong, "Uncertainty-aware breast ultrasound explanation cards: A visual communication framework for image-based AI diagnostic support using BreastMNIST_224," Int. J. Graph. Des., vol. 3, no. 2, pp. 365-380, 2025, doi: 10.51903/ijgd.v3i2.3701.

[38] B. Zhou, C. Li, and L. Liu, "Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication," Int. J. Graph. Des., vol. 3, no. 2, pp. 365-380, 2025, doi: 10.51903/ijgd.v3i2.3696.

[39] C. Li, B. Zhou, and K. Gao, "Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces," Int. J. Graph. Des., vol. 3, no. 2, pp. 381-394, 2025, doi: 10.51903/ijgd.v3i2.3709.

[40] J. Jin, "LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy," Int. J. Graph. Des., vol. 3, no. 2, pp. 397-414, 2025, doi: 10.51903/ijgd.v3i2.3698.

[41] C. Li, J. Bai, and S. Wang, "Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications," J. Adv. Comput. Syst., vol. 4, no. 2, pp. 76-92, 2024, doi: 10.69987/JACS.2024.40207.

[42] Z. S. Zhong, J. Chen, E. Zhong, and X. Sun, "Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 502-520, 2025, doi: 10.51903/jtie.v4i2.536.

[43] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, art. e0118432, 2015, doi: 10.1371/journal.pone.0118432.

[44] S. Lu and D. Zhou, "TinyLLM-assisted intrusion detection for real-time IoT networks," J. Adv. Comput. Syst., vol. 4, no. 8, pp. 72-87, 2024, doi: 10.69987/JACS.2024.40809.

[45] Q. Xin, "Hybrid cloud architecture for efficient and cost-effective large language model deployment," J. Inf. Syst. Informatics, vol. 7, no. 3, pp. 2182-2195, 2025, doi: 10.51519/journalisi.v7i3.1170.

[46] S. He, C. Li, and H. Rao, "Few-shot cold-start workload forecasting for new AI inference tenants with time-series foundation models," J. Technol.

Informatics Eng., vol. 4, no. 1, pp. 306-324, 2025, doi: 10.51903/jtie.v4i1.546.

[47] J. Zhang, "From general human activity recognition to volleyball-oriented wearable transfer learning: Cross-dataset evidence from UCI HAR and WISDM for domain adaptation and edge deployment," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 263-283, 2025, doi: 10.51903/jtie.v4i1.524.

[48] S. He, X. Chang, and E. Sun, "Cross-cloud transfer learning for AI training capacity forecasting under workload and topology distribution shift," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 100-120, 2024, doi: 10.69987/JACS.2024.40108.

[49] Q. Xin, "Probabilistic bike-sharing demand forecasting under changing weather and seasonal regimes with transformer-based models," Findings, Mar. 2026, doi: 10.32866/001c.157499.

[50] S. Chen, S. He, and E. Sun, "Risk-bounded GPU resource oversubscription via conformal demand envelopes in production AI clusters," J. Adv. Comput. Syst., vol. 4, no. 5, pp. 119-134, 2024, doi: 10.69987/JACS.2024.40509.

[51] Q. Xin, "Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation," AVITEC, vol. 8, no. 2, pp. 247-264, 2026, doi: 10.28989/avitec.v8i2.3974.

[52] J. Jin, "Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 520-533, 2025, doi: 10.51903/jtie.v4i2.535.

[53] Q. Xin, "Explainable and fair credit risk scoring with counterfactual explanations: A reproducible evaluation on the German Credit dataset (HELOC-motivated)," J. Inf. Technol., vol. 14, no. 2, pp. 215-231, 2026, doi: 10.32664/j-intech.v14i02.2228.

[54] X. Sun, Z. S. Zhong, and Q. Wu, "Retrieval-grounded HDFS log anomaly detection and deterministic failure narrative generation," J. Comput. Syst. Appl., vol. 3, no. 1, pp. 15-30, 2026, doi: 10.64229/j6d7fr94.

[55] Q. Xin, "Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI Online Retail," J. Inf. Technol., vol. 14, no. 1, pp. 20-37, 2026, doi: 10.32664/j-intech.v14i01.2229.

[56] W. Su, H. Rao, and E. Ma, "Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks," Int. J. Graph. Des., vol. 4, no. 1, pp. 186-191, 2026, doi: 10.51903/ijgd.v4i1.3699.

[57] B. Zhang, H. Rao, and D. Zhao, "Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents," J. Adv. Comput. Syst., vol. 4, no. 3, pp. 109-125, 2024, doi: 10.69987/JACS.2024.40308.

[58] G. Liu, C. Li, and E. Zhang, "OpsLLM for cloud incident triage: Bilingual RAG-based root cause analysis and alert summarization for AI infrastructure operations," J. Adv. Comput. Syst., vol. 4, no. 4, pp. 97-111, 2024, doi: 10.69987/JACS.2024.40408.

[59] W. Su, S. Chen, and C. Zhao, "Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 649-662, 2025, doi: 10.51903/jtie.v4i3.543.

[60] D. Zheng, B. Zhang, and J. Geibel, "VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 67-82, 2024, doi: 10.69987/JACS.2024.40106.

[61] W. Su, S. Chen, and E. Qian, "Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 327-340, 2026, doi: 10.51903/jtie.v5i1.549.

[62] B. Zhou, J. Jin, and D. Zhao, "Calibrated resume-job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 625-648, 2025, doi: 10.51903/jtie.v4i3.529.

[63] Y. Chen, Y. Zhang, and M. Sherman, "Going concern and bankruptcy prediction under extreme class imbalance: Cost-sensitive learning, resampling, and focal loss with explainable financial-ratio portraits," J. Adv. Comput. Syst., vol. 4, no. 4, pp. 80-96, 2024, doi: 10.69987/JACS.2024.40407.

[64] Z. S. Zhong, C. Li, and H. Rao, "Trajectory reliability prediction for generalist AI agents: Tool-use failure analysis and success forecasting on ZClawBench," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 341-360, 2026, doi: 10.51903/jtie.v5i1.539.

[65] Y. Chen and H. Xu, "Trust-calibrated multilingual RAG for humanitarian information platforms:

Empirical evaluation on OMoS-QA for migration information access," Int. J. Graph. Des., vol. 4, no. 1, pp. 141-164, 2026, doi: 10.51903/ijgd.v4i1.3552.

[66] B. Zhou, H. Wang, and X. Chang, "Distilling VMAF into an edge-deployable quality predictor: A pilot shot-level proxy with LLM-ready quality tokens," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 447-463, 2025, doi: 10.51903/jtie.v4i2.522.

[67] C. Li, G. Liu, and Z. Zhao, "Cost-aware LLM-style routing for AIOps log analysis: Log parsing, anomaly detection, fault diagnosis, and incident summarization on LogEval task files," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 91-103, 2026, doi: 10.51903/jtie.v5i2.538.

[68] X. Chang, Y. Lu, and Z. S. Zhong, "Review-grounded explainable recommendation with faithfulness evaluation on Amazon reviews," J. Electr. Eng. Comput. Sci., vol. 11, no. 1, pp. 9-22, 2026, doi: 10.54732/jeecs.v11i1.2.

[69] D. Zheng and C. Li, "Behavior-level jailbreak resistance via multi-stage refusal + utility preservation," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 83-99, 2024, doi: 10.69987/JACS.2024.40107.

[70] Z. Li, K. Zhang, and A. Wong, "Numerical-reasoning guardrails for a quant research assistant: A compact reproducible benchmark using SEC and FRED data," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 75-90, 2026, doi: 10.51903/jtie.v5i2.541.

[71] B. Zhang, X. Sun, G. Liu, and B. Zhou, "LLM-style DevOps copilot for cloud-native troubleshooting: Retrieval-augmented runbook generation and command-safety evaluation," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 104-118, 2026, doi: 10.51903/jtie.v5i2.534.

[72] M.-J. Kuo, D. Zheng, and J. Hires, "Federated topic-preference learning for knowledge-grounded chat with differential privacy," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 385-401, 2025, doi: 10.51903/jtie.v4i2.502.

[73] Y. Zhang and H. Zhang, "A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 464-486, 2025, doi: 10.51903/jtie.v4i2.547.

[74] J. Nie, G. Liu, C. Li, and T. Zou, "Evidence-constrained incident visualization cards for distributed cloud logs: A UI/UX framework for turning Hadoop, OpenStack, and ZooKeeper logs into actionable SRE design interfaces," Int. J. Graph. Des., vol. 4, no. 1, pp. 179-185, 2026, doi: 10.51903/ijgd.v4i1.3703.

[75] Y. Zhang and H. Zhang, "Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces," Int. J. Graph. Des., vol. 3, no. 1, pp. 214-229, 2025, doi: 10.51903/ijgd.v3i1.3722.

[76] Z. Li, S. Zhou, and Z. Zhou, "Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench," Int. J. Graph. Des., vol. 3, no. 1, pp. 196-210, 2025, doi: 10.51903/ijgd.v3i1.3715.

[77] Y. Li, S. Lu, and L. Zhao, "LLM-as-design-critic: Aligning AI-generated UI feedback with human graphic design judgment," Int. J. Graph. Des., vol. 3, no. 1, pp. 196-215, 2025, doi: 10.51903/ijgd.v3i1.3661.