This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available at SpringerLink.

Discrimination Is Not Enough: Calibration-Aware Multimodal Classification of Benign and Malignant Lesions in Contrast-Enhanced Spectral Mammography

Simbiat Damilola Adetoro
Peter Ifeoluwa Adeyemo
Fatima Ez-Zahraa Bazay
Confidence Raymond
Aondona Moses Iorumbur

Abstract

Most breast lesion classifiers prioritize accuracy but overlook whether their confidence can be trusted. For benign-versus-malignant (BvM) classification, an overconfident wrong diagnosis is more dangerous than honest abstention, as well-calibrated confidence allows benign cases to be safely cleared without a clinician, easing workload in low-resource settings. We evaluate calibration safety in multimodal fusion using the CDD-CESM dataset of 326 patients with paired low-energy (LE) and dual-energy subtracted (DES) mammography and structured radiologist text. Clinical text closes a real discrimination gap, as image-only models reach AUC 0.732–0.752 while fusion raises this to 0.837–0.899 ($p < 0.0001$). Calibration effects are mixed, as ECE improves for LE under fusion from 0.203 to 0.127 but degrades for DES from 0.047 to 0.112. At a single fixed threshold, both fusion models appear less safe than image-only baselines. Sweeping all thresholds reverses this for LE, as LE+Text reaches zero missed malignancies at only 85% deferral, fewer than either image-only model (90–92%), while DES+Text remains the least safe configuration under both evaluations, requiring 94.9% deferral for the same guarantee. The implementation is available at: https://github.com/joyinola/BI-RADS-descriptor-aware-mammography-AI-for-safe-referral-triage-in-LMIC-screening

1 Introduction

A breast lesion classifier reporting 90% accuracy has told a clinician little about whether its prediction can be trusted on the next patient. Area Under the Curve (AUC) and similar metrics evaluate average class separation, revealing nothing about whether a model's stated confidence on any individual prediction is meaningful. A model can be highly discriminative yet poorly calibrated, *i.e.* confidently wrong as often as confidently right, a distinction standard mammography AI reporting rarely acknowledges (Aboutalib et al. 2018; Acosta-Jiménez et al. 2025; Helal et al. 2024; Qasrawi et al. 2024). This is most critical for Benign vs. Malignant (BvM) classification, which determines whether a patient proceeds to biopsy: an overconfident wrong diagnosis is far costlier than honest abstention, and where clinician shortages create bottlenecks, calibrated confidence is what lets benign cases be safely cleared, preserving clinician time for ambiguous cases. We evaluate this in Contrast-Enhanced Spectral Mammography (CESM), which provides a low-energy (LE) image equivalent to standard digital mammography and a dual-energy subtracted (DES) image highlighting contrast uptake from

tumour vascularity (Jochelson and Lobbes 2021; Khaled et al. 2022), paired with structured descriptors such as Breast Imaging Reporting and Data System (BI-RADS) score, breast density, and free-text findings. CESM offers near-MRI sensitivity at a fraction of the cost (Xiang et al. 2020; Gelardi et al. 2022) and deep learning is frequently applied to its image streams (Aboutalib et al. 2018; Ribli et al. 2018), but how fusing clinical text affects calibration and deployment safety, not just discrimination, remains unexamined. This is harder than it looks since BI-RADS is highly discriminative: a model relying on it can become more confident without becoming more reliable, and discrimination and calibration can move independently or oppositely. Using the CDD-CESM dataset (Khaled et al. 2022), we restrict our central analysis to BvM, the most challenging boundary (Acosta-Jiménez et al. 2025) and, as Section 3.1 shows, the only one where BI-RADS does not already give the label away. We aim to address the following research questions: **RQ1.** Does fusing structured clinical text into LE and DES image streams improve discrimination on the hardest diagnostic boundary, Benign vs. Malignant? **RQ2.** Does a discrimination gain from text fusion also translate into a calibration gain, or can the two move independently? **RQ3.** Does the safety ranking implied by a single fixed triage threshold hold once the full range of operating points is examined?

Our three contributions are: (1) a late score-fusion framework integrating clinical text with both LE and DES streams, with discrimination gains verified via DeLong's paired test (DeLong et al. 1988); (2) an evaluation pipeline exposing calibration alongside discrimination across four uncertainty quantification (UQ) strategies; and (3) a continuous-threshold triage simulation evaluated across the full spectrum of operational thresholds rather than a single fixed cutoff.

2 Related Work

Deep convolutional architectures, including ResNet (He et al. 2016), DenseNet (Huang et al. 2017), and EfficientNet (Tan and Le 2019), have been applied extensively to mammographic classification, with reported AUC between 0.76 and 0.97 depending on task and modality (Aboutalib et al. 2018; Ribli et al. 2018; Helal et al. 2024; Qasrawi et al. 2024). Acosta-Jiménez *et al.* (Acosta-Jiménez et al. 2025) compared DM and CESM on CDD-CESM under identical preprocessing, training, and evaluation. CESM outperformed DM on Benign vs. Malignant through contrast-driven lesion conspicuity, while DM best separated Normal vs. Malignant (EfficientNet-B0 AUC 0.973 vs. 0.933). SHAP maps aligned to anatomically consistent regions in both modalities despite this quantitative gap.

As image and text streams become increasingly available together, multimodal fusion strategies are gaining importance. Boulahia *et al.* (Boulahia et al. 2021) compare early, intermediate, and late fusion for multimodal action recognition using RGB, depth, and skeletal streams, finding that intermediate fusion with explicit feature selection outperforms naive concatenation, and that a learned late-fusion network over modality-wise scores substantially outperforms handcrafted score-combination rules.

Beyond discrimination, model confidence itself has become an important metric of performance. Monte Carlo Dropout (MCD) (Gal and Ghahramani 2016), Test-Time Augmentation (TTA) (Wang et al. 2019), and Deep Ensembles (Lakshminarayanan et al. 2017) are the standard UQ approaches in deep learning. Maganti (Maganti 2026) studies the related problem of converting uncertainty into deferral decisions for medical image segmentation, reporting that

calibration improvements do not reliably translate into better decision quality; the two properties must be evaluated separately.

3 Methods

3.1 Dataset and Task Scoping

We used the CDD-CESM dataset (Khaled et al. 2022), comprising 2006 CESM examinations from 326 patients, each providing a paired LE and DES image alongside a structured radiologist descriptor (BI-RADS score, breast density, and free-text findings), with every case labelled Normal, Benign, or Malignant.

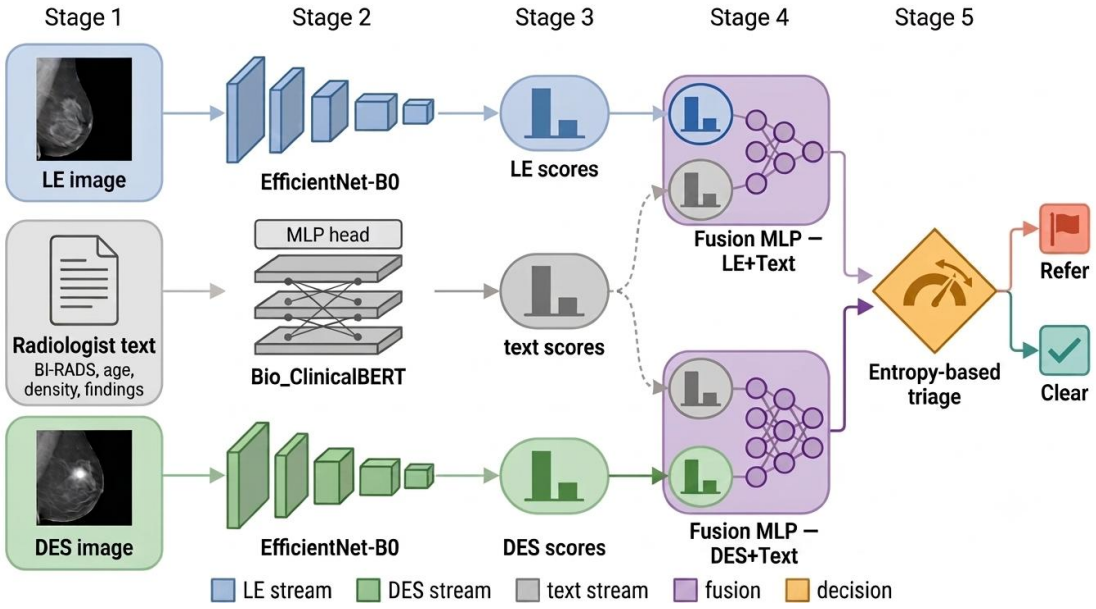
Not every diagnostic boundary here is equally informative. BI-RADS alone already separates Normal from the other two classes almost perfectly: a score of 1 is Normal in 88.7% of cases and a score of 5 is Malignant in 96.9%, so strong accuracy on Normal vs. Benign or Normal vs. Malignant would largely measure whether a model can read a score, not whether it adds value beyond it. Benign vs. Malignant avoids this: every BvM case falls in a BI-RADS bin, predominantly 3 and 4, with substantial cases from the opposite class (BI-RADS 4 is 39 Benign and 155 Malignant), so resolving BvM requires information beyond the score. We therefore restrict our calibration, uncertainty, and triage analysis to BvM, reporting Normal vs. Benign and Normal vs. Malignant only as discrimination benchmarks against the published baseline. Splits were performed at the patient level to prevent exam leakage across training, validation, and test sets.

3.2 Architectures

Image streams. LE and DES images are each classified independently by an EfficientNet-B0 (Tan and Le 2019) backbone, fine-tuned from ImageNet initialization, following the training protocol of the published baseline (Acosta-Jiménez et al. 2025), so our image-only numbers remain directly comparable to it.

Text stream. Each examination's free-text radiologist findings, together with age, BI-RADS score, and breast density encoded as natural-language tokens in the same string, are embedded with a frozen Bio_ClinicalBERT (Alsentzer et al. 2019) encoder. The resulting 768-dimensional embedding passes through a lightweight multilayer perceptron (MLP) trained directly on BvM. We embed structured and free-text fields jointly rather than in separate branches, as with only hundreds of training examples, a frozen language model benefits from clinical scores appearing as context within its pretraining input, rather than requiring a second branch to relearn the same numeric relationships with too little data. We freeze the encoder for the same reason, since fine-tuning a 110M-parameter model on hundreds of cases risks overfitting.

Late fusion. Following Boulahia *et al.*'s (Boulahia et al. 2021) finding that learned score combination outperforms handcrafted fusion rules, we train each image and text stream independently, then combine their predicted probabilities with a small fully connected network taking as input the softmax outputs of both streams and passing them through two hidden layers; the decision threshold for each fusion model is selected on a held-out validation set by maximizing the F1 score (Figure 1).



Benign-versus-malignant classification pipeline. LE and DES images are classified with EfficientNet-B0, while radiologist text (findings, age, BI-RADS, and breast density) is encoded by a frozen Bio_ClinicalBERT and MLP. Separate fusion MLPs combine image and text scores. LE, DES, LE + Text, and DES + Text use the same entropy-based triage rule, referring uncertain or high-risk cases and clearing the rest.

3.3 Evaluation Protocol

Discrimination. We report AUC for every condition and test each fusion model against its corresponding image-only baseline with DeLong's paired test (DeLong et al. 1988) for AUC differences, which accounts for both models being scored on the same test cases.

Calibration. We compute Expected Calibration Error (ECE) (Guo et al. 2017) under four inference regimes, namely standard single-pass inference, Monte Carlo Dropout (MCD, 20 passes), TTA (5 augmentation policies), and a 3-seed Deep Ensemble, with 95% confidence intervals from 2000 bootstrap resamples, as ECE has no closed-form variance estimator; two ECE values are distinguishable only when their intervals do not overlap.

Triage and operating-point sweep. We convert each model's predictions into a three-way decision (High Risk, Low Risk, or Uncertain) by filtering predictive entropy through thresholds optimized on the validation set. Entropy is computed directly from each model's continuous output probability, prior to the F1-tuned decision threshold; that threshold is applied only to cases clearing the entropy gate, splitting them into High Risk or Low Risk. Uncertain cases are deferred to a clinician. We report the deferral rate (Uncertain plus High Risk), and the fraction of malignant cases incorrectly placed in Low Risk, our safety-critical quantity. Rather than a single fixed cutoff, which conflates safety capacity with efficiency, we sweep the entropy cutoff across its full range to find each configuration's "safety floor", defined as the minimum deferral rate for zero missed malignancies. Because several error rates fall at or near zero, we use Wilson score intervals (Wilson 1927) throughout for valid bounds.

4 Results

4.1 Choosing a calibration regime

Addressing RQ2, before committing to a single regime for the safety analysis in Section 4.3, we first check whether the choice of UQ strategy changes the calibration picture (Table 1). No regime is statistically distinguishable from Standard inference for any of the four conditions, as every regime's 95% CI overlaps Standard's CI for the same condition, despite point-estimate differences that can look large in isolation. DES alone is better calibrated than LE alone under Standard inference and MCD, where the two conditions' intervals do not overlap; under TTA and Deep Ensembles, the two intervals overlap and the conditions are comparable. We report Standard inference throughout the remainder of this paper, as it is the regime under which our calibration ranking is supported by non-overlapping intervals. This near-uniform overlap across UQ regimes is consistent with our modest sample size of 71 malignant test cases, where bootstrap ECE intervals are wide enough that even substantial point-estimate shifts from MCD, TTA, or Ensembling rarely separate from Standard inference. We treat this as a property of the evaluation budget rather than evidence that UQ choice is inconsequential at larger scale.

ECE under four uncertainty quantification regimes, Benign vs. Malignant, with 95% bootstrap confidence intervals.

Condition	Standard		MCD		TTA		Deep Ensemble	
LE alone	0.203	[0.150, 0.295]	0.208	[0.149, 0.296]	0.084	[0.050, 0.171]	0.122	[0.084, 0.215]
DES alone	0.047	[0.030, 0.141]	0.047	[0.031, 0.138]	0.061	[0.034, 0.143]	0.147	[0.081, 0.229]
LE + Text	0.127	[0.081, 0.201]	0.129	[0.078, 0.199]	0.055	[0.041, 0.134]	0.094	[0.056, 0.164]
DES + Text	0.112	[0.073, 0.172]	0.120	[0.071, 0.176]	0.116	[0.070, 0.175]	0.102	[0.065, 0.168]

4.2 Clinical text improves discrimination

Addressing RQ1, on the diagnostic boundary where discrimination is genuinely difficult, fusing clinical text into either image stream produces a large, statistically unambiguous improvement (Table 2). Regardless of modality or how safety is subsequently evaluated, the model that sees clinical text is reliably better at distinguishing benign from malignant cases.

Discrimination on Benign vs. Malignant. DeLong p-values compare each fusion model against its own image-only baseline on the same test cases.

Condition	AUC	95% CI (bootstrap)	Δ AUC	p
LE alone	0.732	[0.646, 0.817]	–	–
DES alone	0.752	[0.652, 0.836]	–	–
LE + Text	0.837	[0.768, 0.898]	+0.105	3.4×10^{-7}
DES + Text	0.899	[0.832, 0.952]	+0.147	2.1×10^{-5}

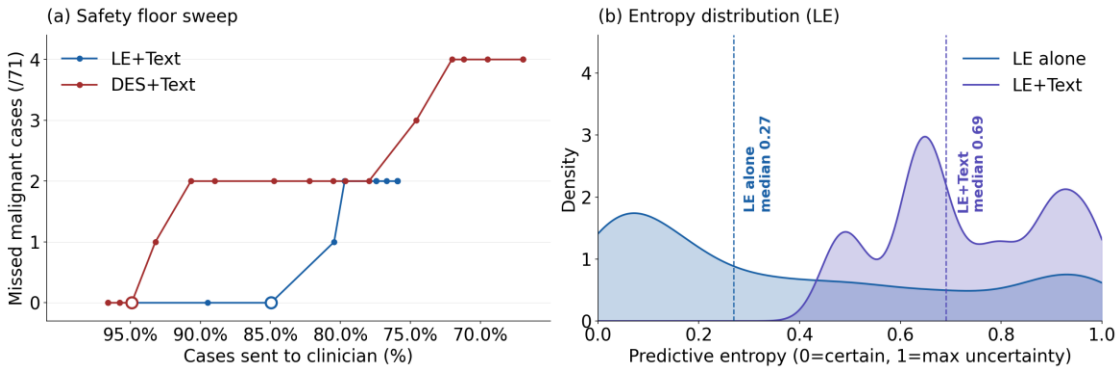
4.3 From one fixed threshold to the full operating-point sweep

When evaluated the conventional way, by fixing a single confidence cutoff and counting missed malignant cases, multimodal fusion appears to introduce a safety trade-off. Out of 71 malignant test cases, both image-only models miss none, while LE + Text misses 2 and DES + Text misses 4 (Table 3); their confidence intervals overlap, so this difference is not statistically significant, but the broader pattern, image-only models are safe and fusion models are not, is exactly what a single-threshold evaluation would report and is the conclusion a less careful study would draw. The same fixed cutoff also establishes a clear calibration ranking with DES alone best-calibrated and LE alone worst, the only pair among six possible comparisons that is statistically distinguishable; both fusion models sit between the two image-only extremes and are not distinguishable from either or from each other.

Safety at a single fixed entropy cutoff versus each model's own safety floor (the lowest deferral rate with zero missed malignant cases), Benign vs. Malignant, Wilson 95% confidence intervals.

	Fixed Cutoff		Safety Floor	
2-3 (lr)4-5 Condition	Missed / 71	Sent to clinician	Missed / 71	Sent to clinician
LE alone	0 (0.0% [0.0, 5.1])	90.2% [84.0, 94.2]	0 (0.0% [0.0, 5.1])	90.2% [84.0, 94.2]
DES alone	0 (0.0% [0.0, 5.1])	92.4% [86.1, 95.9]	0 (0.0% [0.0, 5.1])	92.4% [86.1, 95.9]
LE + Text	2 (2.8% [0.8, 9.7])	75.9% [68.0, 82.4]	0 (0.0% [0.0, 5.1])	85.0% [77.9, 90.0]
DES + Text	4 (5.6% [2.2, 13.6])	72.0% [63.3, 79.3]	0 (0.0% [0.0, 5.1])	94.9% [89.3, 97.6]

Addressing RQ3, a single cutoff is an arbitrary point on a curve, not a property of the model. Sweeping the entropy cutoff and recording, at each point, the deferral rate and missed malignant cases (Figure 2a) shows that the ranking above does not survive once every model is evaluated at its own safety floor rather than a single shared default. LE alone reaches zero missed cases only at 90.2% deferral; DES alone needs 92.4%. LE + Text reaches the same guarantee at only 85.0%, lower than either image-only model, the most efficient safe configuration we test and the opposite of its ranking at the shared default cutoff. DES + Text does not improve the same way, reaching zero missed cases only at 94.9% deferral, and that guarantee is fragile, as a single additional case appears once deferral drops below 93.2%, never returning to zero as the cutoff relaxes further.



(a) Safety floor analysis for LE + Text and DES + Text on Benign vs. Malignant. Open circles denote the safety floor, the lowest deferral rate with zero missed malignant cases. LE-image and DES-image are omitted as they achieve zero misses throughout. The x-axis decreases from left to right. (b) Predictive entropy distributions for LE and LE + Text. Dashed lines mark median entropy; text fusion shifts the median from 0.27 to 0.69, consistent with the improved safety floor in (a).

5 Discussion

The central lesson is not any single number in Section 4 but that three reasonable ways of checking the same four models give three different answers: on discrimination fusion wins clearly for both modalities; at one fixed threshold fusion looks like the clear loser for both; across the full range of thresholds fusion is the better choice for LE and comparable for DES. None of these checks is wrong, but a study reporting only one will reach a confidently incorrect conclusion about the others. The failure mode is asymmetric: a model whose apparent safety comes from being uncertain about nearly every case looks identical to a genuinely well-calibrated one when safety is checked at a single cutoff, since both report zero misses, and only the operating-point sweep separates them, consistent with calibration gains not reliably tracking decision quality (Maganti 2026; Mehrtash et al. 2020).

LE and DES alone are unsafe for opposite reasons: LE is confident (median entropy 0.27) yet miscalibrated (ECE 0.203), forcing a strict cutoff, while DES's strong ECE of 0.047 is hollow because it hedges on nearly every case (median entropy 0.91). Fusion reverses both, pulling LE toward more hedging (entropy 0.69), which lowers its floor to 85.0%, and DES toward sharper commitments (entropy 0.76) at the cost of a higher floor of 94.9%. This matters for resource-constrained settings, where the practical value of a triage tool is how much clinician time it saves while remaining safe, not whether it can be made safe at all. LE + Text reaches the zero-missed-malignancy guarantee while sending only 85% for review, clearing roughly 53% more cases than LE alone and nearly double DES alone, at negligible inference cost (two EfficientNet-B0 passes and one frozen Bio_ClinicalBERT embedding, feasible on CPU). The right comparison is thus not which condition has the highest AUC, but which reaches the required safety floor at the lowest deferral rate.

6 Conclusion

Text fusion reliably improves discrimination for both CESM modalities on BvM, but its effect on safety is modality-dependent and only visible across the full range of triage operating points. Future work should extend this evaluation to larger, more diverse cohorts and examine whether the entropy-based triage rule generalises across scanner types and patient populations beyond the single-site CDD-CESM dataset used here. A model's discrimination, calibration, and deployment safety do not move together, as checking only one, or only a single fixed threshold, can produce a confidently wrong deployment decision. Our sample size for the safety-critical comparisons is modest at 71 malignant test cases; the ranking between LE + Text and DES + Text on missed-malignancy rate is not statistically distinguishable at this size even though point estimates differ, and we report Wilson intervals throughout so this is visible rather than implied.

7 Limitations and Future Work

Our evaluation draws on a single-site cohort with a modest number of malignant test cases, so the reported estimates may not transfer to other scanners, populations, or acquisition protocols. In addition, the text stream consumes radiologist-authored descriptors (BI-RADS, breast density, free-text findings) that are available for every case in CDD-CESM, so the deployment value we report assumes text is present at triage. Where such descriptors are unavailable at inference, fusion becomes a missing-modality problem and the reported safety floor may not hold. We leave evaluation on larger, cross-acquisition cohorts and the characterization of performance under partial or absent text to future work.

8 Acknowledgeament

The authors would like to thank the following instructors of the Sprint AI Training for African Medical Imaging Knowledge Translation (SPARK) Academy 2026 summer school on deep learning in medical imaging for providing insightful background knowledge on tumours that informed the research presented here. The authors would also like to thank Linshan Liu for administrative assistance in supporting the SPARK Academy training and capacity-building activities, which the authors immensely benefited from. The authors acknowledge the computational infrastructure support from the Digital Research Alliance of Canada (The Alliance) and knowledge translation support from the McGill University Doctoral Internship program through the student exchange program for the SPARK Academy. Finally, we would like to thank Lacuna Fund for Health and Equity, the Radiological Society of North America (RSNA), Research and Education (R&E) Foundation Derek Harwood-Nash International Education Scholar Grant, and National Science and Engineering Research Council of Canada (NSERC) Discovery Launch Supplement for making the SPARK Academy possible via research grant supports.

Aboutalib, Sarah S., Aly A. Mohamed, Wendie A. Berg, Margarita L. Zuley, Jules H. Sumkin, and Shandong Wu. 2018. "Deep Learning to Distinguish Recalled but Benign Mammography Images in Breast Cancer Screening." *Clin. Cancer Res.* 24 (23): 5902–9. https://doi.org/10.1158/1078-0432.CCR-18-1115.

Acosta-Jiménez, Samara, Miguel M. Mendoza-Mendoza, Carlos E. Galván-Tejada, José M. Celaya-Padilla, Jorge I. Galván-Tejada, and Manuel A. Soto-Murillo. 2025. "Explainable Deep Learning for Breast Lesion Classification in Digital and Contrast-Enhanced Mammography." *Diagnostics* 15 (24): 3143. https://doi.org/10.3390/diagnostics15243143.

Alsentzer, Emily, John R. Murphy, Willie Boag, et al. 2019. "Publicly Available Clinical BERT Embeddings." *Proc. 2nd Clinical NLP Workshop*, 72–78. https://doi.org/10.18653/v1/W19-1909.

Boulahia, Said Yacine, Abdenour Amamra, Mohamed Ridha Madi, and Said Daikh. 2021. "Early, Intermediate and Late Fusion Strategies for Robust Deep Learning-Based Multimodal Action Recognition." *Mach. Vision Appl.* 32 (6): 121. https://doi.org/10.1007/s00138-021-01249-8.

DeLong, Elizabeth R., David M. DeLong, and Daniel L. Clarke-Pearson. 1988. "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach." *Biometrics* 44 (3): 837–45. https://doi.org/10.2307/2531595.

Gal, Yarin, and Zoubin Ghahramani. 2016. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." *International Conference on Machine Learning*, 1050–59.

Gelardi, Fabrizia, Eleonora M. Ragaini, Martina Sollini, Daniela Bernardi, and Arturo Chiti. 2022. "Contrast-Enhanced Mammography Versus Breast Magnetic Resonance Imaging: A Systematic Review and Meta-Analysis." *Diagnostics* 12 (8): 1890. https://doi.org/10.3390/diagnostics12081890.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. "On Calibration of Modern Neural Networks." *International Conference on Machine Learning*, 1321–30.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep Residual Learning for Image Recognition." *IEEE Conference on Computer Vision and Pattern Recognition*, 770–78. https://doi.org/10.1109/CVPR.2016.90.

Helal, Mohamed, Rana Khaled, Omar Alfarghaly, Omar Mokhtar, Abeer Elkorany, and Ayman Aly. 2024. "Validation of Artificial Intelligence Contrast Mammography in Diagnosis of Breast Cancer: Relationship to Histopathological Results." *Eur. J. Radiol.* 173: 111392. https://doi.org/10.1016/j.ejrad.2024.111392.

Huang, Gao, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. "Densely Connected Convolutional Networks." *IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708. https://doi.org/10.1109/CVPR.2017.243.

Jochelson, Maxine S., and Marc B. I. Lobbes. 2021. "Contrast-Enhanced Mammography: State of the Art." *Radiology* 299 (1): 36–48. https://doi.org/10.1148/radiol.2021201948.

Khaled, Rana, Maha Helal, Omar Alfarghaly, et al. 2022. "Categorized Contrast Enhanced Mammography Dataset for Diagnostic and Artificial Intelligence Research." *Sci. Data* 9: 122. https://doi.org/10.1038/s41597-022-01238-0.

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2017. “Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles.” *Advances in Neural Information Processing Systems* 30.

Maganti, Saket. 2026. “Rethinking Uncertainty in Segmentation: From Estimation to Decision.” *arXiv Preprint arXiv:2604.13262*.

Mehrtash, Alireza, William M. Wells, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur. 2020. “Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation.” *IEEE Transactions on Medical Imaging* 39 (12): 3868–78. https://doi.org/10.1109/TMI.2020.3006437.

Qasrawi, Radwan, Omar Daraghmeh, Ibrahim Qdaih, et al. 2024. “Hybrid Ensemble Deep Learning Model for Advancing Breast Cancer Detection and Classification in Clinical Applications.” *Heliyon* 10: e38374. https://doi.org/10.1016/j.heliyon.2024.e38374.

Ribli, Dezső, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. 2018. “Detecting and Classifying Lesions in Mammograms with Deep Learning.” *Sci. Rep.* 8: 4165. https://doi.org/10.1038/s41598-018-22437-z.

Tan, Mingxing, and Quoc V. Le. 2019. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.” *International Conference on Machine Learning*, 6105–14.

Wang, Guotai, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. 2019. “Aleatoric Uncertainty Estimation with Test-Time Augmentation for Medical Image Segmentation with Convolutional Neural Networks.” *Neurocomputing* 338: 34–45. https://doi.org/10.1016/j.neucom.2019.01.103.

Wilson, Edwin B. 1927. “Probable Inference, the Law of Succession, and Statistical Inference.” *Journal of the American Statistical Association* 22 (158): 209–12. https://doi.org/10.2307/2276774.

Xiang, Wenrui, Hao Rao, and Lihua Zhou. 2020. “A Meta-Analysis of Contrast-Enhanced Spectral Mammography Versus MRI in the Diagnosis of Breast Cancer.” *Thorac. Cancer* 11 (6): 1423–32. https://doi.org/10.1111/1759-7714.13400.