One Pixel Can Change the Diagnosis: Adversarial and Non-Adversarial Robustness and Uncertainty in Breast Ultrasound Classification Model

Kuan Huang[1]*, **Noorul Sahel**[1], **Dikshya Karki**[2], **Meng Xu**[1], **Yingfeng Wang**[2]*

[1]Department of Computer Science and Technology, Kean University, Union, NJ, United States
[2]Department of Computer Science and Engineering, University of Tennessee at Chattanooga, Chattanooga, TN, United States
{khuang, saheln, mexu}@kean.edu, xhl358@mocs.utc.edu, yingfeng-wang@utc.edu

Abstract

Deep learning models have strong potential for automating breast ultrasound (BUS) image classification to support early cancer detection. However, their vulnerability to small input perturbations poses a challenge for clinical reliability. This study examines how minimal pixel-level changes affect classification performance and predictive uncertainty, using the BUSI dataset and a ResNet-50 classifier. Two perturbation types are evaluated: (1) adversarial perturbations via the One Pixel Attack and (2) non-adversarial, device-related noise simulated by setting a single pixel to black. Robustness is assessed alongside uncertainty estimation using Monte Carlo Dropout, with metrics including Expected Kullback–Leibler divergence (EKL), Predictive Variance (PV), and Mutual Information (MI) for epistemic uncertainty, and Maximum Class Probability (MP) for aleatoric uncertainty. Both perturbations reduced accuracy, producing 17 and 29 "fooled" test samples, defined as cases classified correctly before but incorrectly after perturbation, for the adversarial and non-adversarial settings, respectively. Samples that remained correct are referred to as "unfooled." Across all metrics, uncertainty increased after perturbation for both groups, and fooled samples had higher uncertainty than unfooled samples even before perturbation. We also identify spatially localized "uncertainty-decreasing" regions, where individual single-pixel blackouts both flipped predictions and reduced uncertainty, creating overconfident errors. These regions represent high-risk vulnerabilities that could be exploited in adversarial attacks or addressed through targeted robustness training and uncertainty-aware safeguards. Overall, combining perturbation analysis with uncertainty quantification provides valuable insights into model weaknesses and can inform the design of safer, more reliable AI systems for BUS diagnosis.

Code — https://github.com/kuanhuang0624/one-pixel-bus

Introduction

Breast cancer remains one of the leading causes of cancer-related mortality among women worldwide (Siegel et al. 2025), and early detection is essential for improving patient outcomes (Dan et al. 2024). Breast ultrasound (BUS) is widely used for breast cancer screening and diagnosis due to its non-invasive nature, cost-efficiency, absence of ionizing radiation, and suitability for dense breast tissue (Dan et al. 2024; Afrin et al. 2023). In recent years, convolutional neural networks (CNNs) and other deep learning methods have demonstrated remarkable performance in classifying BUS images, achieving accuracies comparable to or even exceeding those of human experts (Habib et al. 2020; Dan et al. 2024; Huang, Xu, and Qi 2021). Typically, BUS image classification involves categorizing images into benign, malignant, or normal classes, or assigning categories based on BI-RADS scores (Xu et al. 2024). In (Rodriguez, Huang, and Xu 2024), three prominent image classification architectures, including convolutional neural networks (ResNet (He et al. 2016)), transformers (Swin Transformer (Liu et al. 2021)), and Vmamba (Liu et al. 2024), were benchmarked on BUS image classification and achieved strong performance. Similarly, He et al. (He et al. 2024) proposed a vision transformer combined with wavelet transformation for BUS image classification, further highlighting the effectiveness of transformer-based approaches in this domain. However, deploying these systems in clinical environments requires greater attention to robustness and reliability under variable and potentially adverse conditions.

Deep learning models are vulnerable to small input perturbations, often imperceptible, that can lead to drastic misclassifications. In natural image domains, even a change in a single pixel can fool strong classifiers, as demonstrated by the One-Pixel Attack (Su, Vargas, and Sakurai 2019) and other adversarial example studies (Goodfellow, Shlens, and Szegedy 2014; Szegedy et al. 2013). In the medical imaging domain, research has shown that adversarial attacks can compromise diagnostic systems across various modalities, including chest X-ray (Asgari Taghanaki, Das, and Hamarneh 2018), CT (Mirsky et al. 2019), and multiple clinical applications (Finlayson et al. 2019). While adversarial robustness is an active research area in general computer vision, its implications in medical imaging, and particularly in BUS classification, remain underexplored. Furthermore, non-adversarial disturbances such as device-induced noise or data corruption may also degrade performance, yet their effect on model confidence has not been systematically studied in BUS.

Beyond achieving high accuracy, **uncertainty quantification (UQ)** is essential in medical AI to ensure safe

*Corresponding author.

decision-making and enable effective human–AI collaboration. Methods such as Monte Carlo (MC) Dropout approximate Bayesian inference to provide uncertainty estimates, allowing the detection of predictions that may be unreliable (Gal and Ghahramani 2016). In the context of model robustness, even minimal perturbations such as single-pixel attacks can drastically alter classification outcomes in natural images (Su, Vargas, and Sakurai 2019), and similar vulnerabilities have been observed in medical imaging domains (Korpihalkola et al. 2021; Sipola and Kokkonen 2021). These perturbations, whether adversarial or naturally occurring (e.g., sensor noise or data corruption), can degrade predictive accuracy and model calibration. Incorporating UQ into perturbation analysis provides a complementary perspective, revealing both when predictions change and how the model's confidence is affected. While prior reviews have emphasized the role of UQ in improving trustworthiness and calibration in medical imaging (Kurz et al. 2022; Seoni et al. 2023), there has been little work systematically combining robustness testing with uncertainty analysis in BUS classification. Such a joint investigation can uncover high-risk failure modes, including cases where minimal changes produce overconfident misclassifications. Identifying these vulnerabilities can inform targeted robustness training, guide model auditing, and support the development of interpretability tools that highlight regions susceptible to harmful perturbations. This integrated approach is particularly relevant in safety-critical domains like medical diagnosis, where the consequences of confident yet incorrect predictions can be severe.

In this work, we present the first systematic investigation of pixel-level perturbations, considering both adversarial perturbations (One-Pixel Attack) and non-adversarial perturbations (One-Pixel Blackout), and examine their combined effects on classification performance and predictive uncertainty in BUS models. We fine-tune a ResNet-50 classifier on the BUSI dataset (Al-Dhabyani et al. 2020) and evaluate uncertainty using multiple metrics: Expected Kullback–Leibler divergence (EKL), Predictive Variance (PV), and Mutual Information (MI) for epistemic uncertainty, as well as the complement of Maximum Class Probability (MP) for aleatoric uncertainty. Our key findings are as follows:

- Minimal single-pixel perturbations can cause substantial drops in classification accuracy, even when only a single pixel is altered.
- For both types of perturbations, uncertainty consistently increases after perturbation, even for samples where the classification remains correct. This suggests that uncertainty metrics can serve as indicators of corrupted or unstable inputs.
- We identify spatially localized "uncertainty-decreasing" regions, where blacking out individual pixels can both change the predicted class and reduce the model's estimated uncertainty, leading to overconfident errors. These regions reveal high-risk vulnerabilities in the model.

These findings have practical implications: high-risk regions could be used to design robustness-enhancing training strategies, guide model auditing, or generate clinical overlays that warn about potentially misleading predictions. Overall, our study integrates robustness analysis, uncertainty quantification, and medical imaging, aiming to improve the reliability of AI-assisted BUS diagnosis.

Materials and Methods

Dataset

We conducted experiments using the BUSI dataset (Al-Dhabyani et al. 2020), a publicly available breast ultrasound image collection designed for breast cancer analysis. The dataset consists of 780 grayscale ultrasound images, comprising 437 benign cases, 210 malignant cases, and 133 normal cases (images without tumors). We randomly divided the dataset into 80% for training and 20% for testing and performed a three-class classification task to categorize BUS images as benign, malignant, or normal.

Implementation Details and Metrics

All experiments were conducted using PyTorch 2.5.0 on a Dell Precision 5820 Workstation running Ubuntu 20.04. The workstation is equipped with an Intel Xeon W-2223 CPU (3.60 GHz), 32 GB of RAM, one NVIDIA RTX 4000 Ada GPU with 20 GB of memory, and two NVIDIA RTX 2000 GPUs with 6 GB of memory each. For breast ultrasound image classification model training, images were resized to 224×224 and normalized using ImageNet statistics (Deng et al. 2009). The model was trained for 10 epochs with a batch size of 32 using the Adam optimizer with a learning rate of 1×10^{-4} and cross-entropy loss. For each image, predictions and uncertainty metrics were computed before and after the perturbations. Classification performance was assessed using accuracy. Uncertainty was measured using Monte Carlo Dropout with 1,000 stochastic forward passes, applying four metrics: Expected Kullback–Leibler divergence (EKL), Mutual Information (MI), Predictive Variance (PV), and Maximum class Probability (MP).

Methods

We developed a systematic framework to assess both adversarial robustness and uncertainty sensitivity of a BUS image classification model under minimal input changes. The method consists of three main components: baseline model training, perturbation generation, and uncertainty quantification.

Baseline Classification Model. A ResNet-50 model pretrained on ImageNet was fine-tuned on the BUSI dataset to classify images into three categories: normal, benign, and malignant. Input images were resized to 224×224 and normalized using ImageNet statistics (Deng et al. 2009). The model was trained using the Adam optimizer with a learning rate of 1×10^{-4}, batch size of 32, and cross-entropy loss for 10 epochs. To support uncertainty estimation, Monte Carlo (MC) Dropout with a dropout rate of 0.5 was applied during both training and inference.

Condition	Train Accuracy	Test Accuracy	Fooled Images (Test)
Clean (Before Attack)	0.9984	0.8846	–
One Pixel Attack	0.9920	0.7756	17
One Pixel Blackout	0.9840	0.6987	29

Table 1: Classification accuracy on the BUSI dataset before and after perturbations. Train and test sets are evaluated under clean images, One Pixel Attack (adversarial), and One Pixel Blackout (non-adversarial).

Pixel-Level Perturbations. To analyze model robustness, we applied two types of minimal changes to images in the BUSI dataset:

- *Adversarial Attack (One Pixel Attack) (Su, Vargas, and Sakurai 2019)*: A single-pixel adversarial attack was performed using the `torchattacks` (Kim 2020) library with default settings, specifying only one modified pixel (`pixels = 1`). For each image, we first ensured the model produced the correct prediction before applying the attack. If the prediction changed to an incorrect class after the attack, the image was marked as *fooled*; otherwise, it was considered *unfooled*.
- *Non-Adversarial Perturbation (One Pixel Blackout)*: To simulate device-level or acquisition noise and analyze the effects of non-adversarial perturbations, we systematically applied a single-pixel blackout by iteratively setting each pixel in the image to black (intensity value of zero) and observing the resulting classification. If any pixel caused a misclassification, the image was labeled as *fooled*. Otherwise, a randomly selected pixel was blacked out, and the image was recorded as *unfooled* to maintain consistency in downstream analysis.

Uncertainty Quantification. We estimated predictive uncertainty using 1000 stochastic forward passes with MC Dropout. Following the methodology established in our prior work (Huang, Xu, and Wang 2025), we computed the following metrics to assess different aspects of uncertainty:

- **Expected Kullback-Leibler Divergence (EKL)**, **Mutual Information (MI)**, and **Predictive Variance (PV)**: Capture epistemic uncertainty by quantifying disagreement across stochastic predictions.
- **Maximum class Probability (MP)**: Reflects aleatoric uncertainty; lower values indicate higher predictive uncertainty. For consistency with other uncertainty metrics in this work, we report the complement of MP (i.e., $1 - \mathrm{MP}$) so that lower values consistently correspond to lower uncertainty across all metrics. The abbreviation "MP" is retained in tables and figures for brevity.

Uncertainty metrics were calculated separately for *fooled* and *unfooled* samples, both before and after adversarial or non-adversarial perturbations, to analyze how uncertainty responds to subtle input variations.

Results

Classification Accuracy. Table 1 summarizes model performance on the BUSI dataset under clean and perturbed conditions. The model achieved high accuracy on clean data, with 99.84% on the training set and 88.46% on the test set. Under the One Pixel Attack, the test accuracy dropped to 77.56%, with 17 test images that were correctly classified before perturbation becoming incorrectly classified afterward. For the One Pixel Blackout condition, test accuracy decreased further to 69.87%, with 29 test images exhibiting the same transition from correct to incorrect classification after perturbation.

Metric	Before	After
EKL	0.001857 ± 0.001094	0.002201 ± 0.001445
MI	0.001837 ± 0.001063	0.002162 ± 0.001344
PV	0.000463 ± 0.000273	0.000549 ± 0.000363
MP	0.296530 ± 0.146526	0.389859 ± 0.172575

Table 2: Aggregate uncertainty metrics (mean $\pm$ std) on **fooled test images** under the One Pixel Attack. Values are computed using 1000 MC-Dropout passes.

Metric	Before	After
EKL	0.000237 ± 0.000763	0.000390 ± 0.000814
MI	0.000245 ± 0.000799	0.000398 ± 0.000834
PV	0.000060 ± 0.000205	0.000097 ± 0.000201
MP	0.048966 ± 0.111573	0.076138 ± 0.118578

Table 3: Aggregate uncertainty metrics (mean $\pm$ std) on **unfooled test images** under the One Pixel Attack. Values are computed using 1000 MC-Dropout passes.

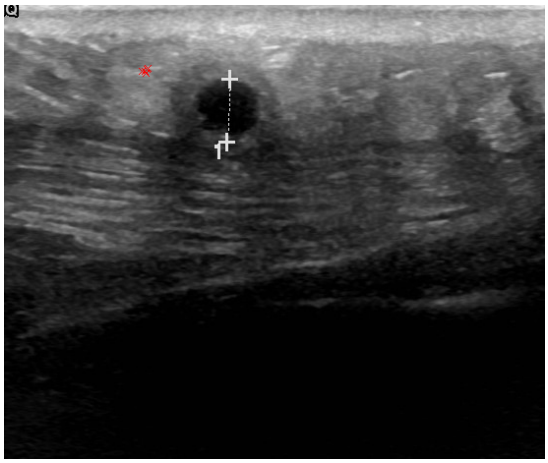
Uncertainty Under One Pixel Attack. Tables 2 and 3 show that for fooled test samples, EKL, MI, PV, and MP all increased after perturbation. For unfooled test samples, all metrics also increased, but with smaller changes and lower overall values compared to fooled test samples.

Uncertainty Under One Pixel Blackout. Tables 4 and 5 indicate that for fooled test samples, all uncertainty metrics increased after perturbation. For unfooled test samples, changes were minimal but consistently positive, and values remained lower than those for fooled test samples.

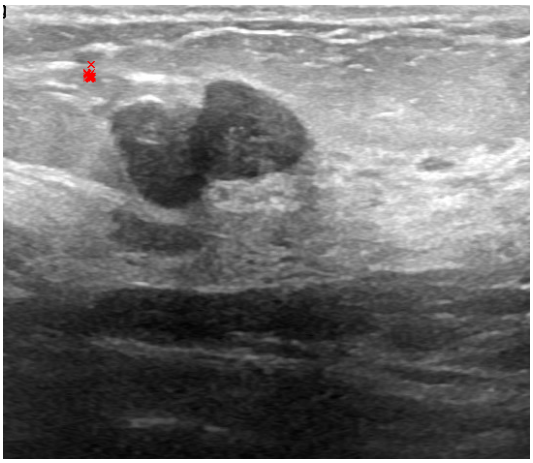
Metric	Before	After
EKL	0.001097 ± 0.001186	0.002924 ± 0.001307
MI	0.001117 ± 0.001231	0.002894 ± 0.001300
PV	0.000275 ± 0.000307	0.000722 ± 0.000319
MP	0.185923 ± 0.174404	0.438953 ± 0.104250

Table 4: Aggregate uncertainty metrics (mean $\pm$ std) on **fooled test images** under the One Pixel Blackout Attack. Values are computed using 1000 MC-Dropout passes.

Uncertainty-Decreasing Single-Pixel Blackouts. In two training images, we identified multiple distinct single pix-



(a)



(b)

Figure 1: Uncertainty-decreasing single-pixel blackouts in two BUS training images. Each red X marks a pixel that, when set to black alone, flips the predicted class and reduces uncertainty compared with the clean image (lower EKL, MI, PV, and MP). Other misclassifying pixels that do not reduce uncertainty are not shown. Subfigures (a) and (b) correspond to two different BUS training images.

Metric	Before	After
EKL	0.000052 ± 0.000181	0.000062 ± 0.000221
MI	0.000054 ± 0.000191	0.000061 ± 0.000218
PV	0.000013 ± 0.000046	0.000015 ± 0.000052
MP	0.016558 ± 0.042479	0.018048 ± 0.045855

Table 5: Aggregate uncertainty metrics (mean $\pm$ std) on **unfooled test images** under the One Pixel Blackout Attack. Values are computed using 1000 MC-Dropout passes.

els where blacking out each pixel *individually* caused the model to change its predicted class. Among these, a subset also reduced the reported MP value, along with decreases in the epistemic metrics EKL, MI, and PV, indicating lower overall uncertainty. Figure 1 marks only these uncertainty-decreasing pixels with red X overlays. This observation shows that very small, localized changes can both alter the classification outcome and suppress stochastic disagreement, resulting in overconfident misclassifications.

Discussion

The results in Table 1 highlight the pronounced vulnerability of the BUS classification model to minimal pixel-level perturbations. While the baseline model achieved strong performance on clean images (99.84% training accuracy and 88.46% test accuracy), a single-pixel modification was sufficient to degrade performance substantially. Under the adversarial One Pixel Attack, test accuracy dropped by 10.9 percentage points, with 17 test images that were correctly classified before perturbation becoming incorrectly classified afterward (fooled samples). The non-adversarial One Pixel Blackout produced an even larger degradation of 18.6 percentage points, with 29 test images exhibiting the same transition from correct to incorrect after perturbation. This finding implies that the model relies on fragile, highly localized features that either adversarial or device-level perturbations can disrupt. The greater drop observed in the blackout setting may be due to the indiscriminate nature of pixel removal, which can eliminate diagnostically relevant structures without gradient-based optimization. Overall, these results underscore the need for robustness mechanisms to mitigate the disproportionate impact of such localized disruptions.

The uncertainty analysis in Tables 2–5 shows that minimal pixel-level perturbations, whether from the One Pixel Attack or the One Pixel Blackout, lead to increased uncertainty in both fooled and unfooled test samples. Across all four metrics (EKL, MI, PV, and MP), values consistently rose after perturbation, indicating that such changes disrupt the model's predictive stability regardless of whether they alter the final classification. This consistent increase across perturbation types suggests that uncertainty estimation can capture the impact of both adversarial and non-adversarial pixel-level changes, highlighting its potential as a tool for detecting corrupted or unstable inputs.

A key observation across both perturbation types is the clear separation in uncertainty values between fooled and unfooled test samples. Before perturbation, fooled test images already exhibited higher uncertainty than unfooled test images, and this gap became larger after perturbation. This pattern suggests that the proposed uncertainty metrics (EKL, MI, PV, and MP) are sensitive to label-changing perturbations, whether they are adversarial or non-adversarial. Such a distinction is important for developing reliability checks in clinical AI systems, as it indicates the potential for uncertainty measures to identify inputs that are at higher risk of being corrupted or adversarially manipulated.

The uncertainty-decreasing single-pixel blackout analysis (Figure 1) shows that, in two training images, multiple pixels within a localized region share a common property: blacking out each pixel individually changes the classification re-

sult and simultaneously reduces the model's predictive uncertainty. The spatial proximity of these pixels indicates that certain regions in the image are particularly sensitive, where even minimal localized modifications can both flip the decision and suppress epistemic and aleatoric uncertainty estimates. Although the overall average uncertainty after perturbation is greater than before, the presence of samples where accuracy decreases while uncertainty also decreases suggests that certain inputs are inherently more susceptible to overconfident misclassification when attacked. This behavior leads to overconfident errors, which are especially concerning in clinical contexts. Identifying these regions could be valuable for targeted model auditing, interpretability analysis, and developing robustness-enhancing strategies, such as focused adversarial training or uncertainty-aware regularization, to reduce the risks associated with localized overconfident errors.

Conclusion and Future Works

This study investigated the impact of minimal pixel-level perturbations on the performance and predictive uncertainty of a breast ultrasound (BUS) image classification model. Using both adversarial (One Pixel Attack) and non-adversarial (One Pixel Blackout) perturbations, we demonstrated that even a single-pixel change can substantially degrade classification accuracy. Uncertainty analysis revealed a clear separation between fooled and unfooled samples: fooled cases consistently exhibited higher uncertainty than unfooled cases, even before perturbation, and this gap widened after perturbation. This indicates that uncertainty estimation can serve as a valuable signal for detecting inputs at elevated risk of misclassification.

We also identified spatially localized regions where multiple single-pixel modifications can individually flip predictions and simultaneously reduce uncertainty, producing overconfident errors. These regions represent potential high-risk areas in the image that can be systematically discovered through perturbation testing and uncertainty analysis. Such regions could be exploited to mount targeted attacks that induce confident misclassifications, but they also offer an opportunity for defense: incorporating these high-risk locations into robustness training or model calibration strategies could strengthen the model against both adversarial and non-adversarial perturbations.

Future work will focus on developing robustness-enhancing strategies such as adversarial training, detection mechanisms, and redundant imaging (e.g., using multiple scans or imaging modalities of the same case so that errors in one image can be corrected or verified by others). Beyond targeted adversarial training on uncertainty-decreasing perturbations, we plan to integrate uncertainty-aware loss functions to improve calibration and leverage localized sensitivity maps to guide model regularization. To strengthen evaluation, we aim to go beyond overall accuracy by reporting metrics such as AUROC, macro-F1, and per-class performance to address class imbalance. Expanding experiments to larger and multi-institutional datasets, as well as exploring alternative training strategies (e.g., early stopping and longer training schedules), will help assess generalizability and reduce overfitting risks. Finally, incorporating interpretability tools (e.g., saliency maps and heatmaps) to explain why specific pixels or regions are critical will enhance clinical trust. Ultimately, our goal is to integrate robustness and uncertainty quantification into practical clinical AI pipelines, ensuring safer and more reliable deployment in real-world diagnostic settings.

Acknowledgments

This work was performed with partial support from the National Science Foundation under Grants Nos. 2430746 and 2430747. This work was also partially supported by the Office of Research and Sponsored Programs, Kean University.

References

Afrin, H.; Larson, N. B.; Fatemi, M.; and Alizad, A. 2023. Deep learning in different ultrasound methods for breast cancer, from diagnosis to prognosis: current trends, challenges, and an analysis. *Cancers*, 15(12): 3139.

Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; and Fahmy, A. 2020. Dataset of breast ultrasound images. *Data in brief*, 28: 104863.

Asgari Taghanaki, S.; Das, A.; and Hamarneh, G. 2018. Vulnerability analysis of chest X-ray image classification against adversarial attacks. In *International Workshop on Machine Learning in Clinical Neuroimaging*, 87–94. Springer.

Dan, Q.; Xu, Z.; Burrows, H.; Bissram, J.; Stringer, J. S.; and Li, Y. 2024. Diagnostic performance of deep learning in ultrasound diagnosis of breast cancer: a systematic review. *NPJ Precision Oncology*, 8(1): 21.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Finlayson, S. G.; Bowers, J. D.; Ito, J.; Zittrain, J. L.; Beam, A. L.; and Kohane, I. S. 2019. Adversarial attacks on medical machine learning. *Science*, 363(6433): 1287–1289.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Habib, G.; Kiryati, N.; Sklair-Levy, M.; Shalmon, A.; Halshtok Neiman, O.; Faermann Weidenfeld, R.; Yagil, Y.; Konen, E.; and Mayer, A. 2020. Automatic breast lesion classification by joint neural analysis of mammography and ultrasound. In *Workshop on Clinical Image-Based Procedures*, 125–135. Springer.

He, C.; Diao, Y.; Ma, X.; Yu, S.; He, X.; Mao, G.; Wei, X.; Zhang, Y.; and Zhao, Y. 2024. A vision transformer network with wavelet-based features for breast ultrasound classification. *Image Analysis and Stereology*, 43(2): 185–194.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, K.; Xu, M.; and Qi, X. 2021. NGMMs: Neutrosophic Gaussian mixture models for breast ultrasound image classification. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 3943–3947. IEEE.

Huang, K.; Xu, M.; and Wang, Y. 2025. Using Adversarial Training to Improve Uncertainty Quantification. *IEEE Transactions on Artificial Intelligence*.

Kim, H. 2020. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*.

Korpihalkola, J.; Sipola, T.; Puuska, S.; and Kokkonen, T. 2021. One-pixel attack deceives computer-assisted diagnosis of cancer. In *Proceedings of the 2021 4th International Conference on Signal Processing and Machine Learning*, 100–106.

Kurz, A.; Hauser, K.; Mehrtens, H. A.; Krieghoff-Henning, E.; Hekler, A.; Kather, J. N.; Fröhling, S.; Von Kalle, C.; Brinker, T. J.; et al. 2022. Uncertainty estimation in medical image classification: systematic review. *JMIR Medical Informatics*, 10(8): e36427.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37: 103031–103063.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Mirsky, Y.; Mahler, T.; Shelef, I.; and Elovici, Y. 2019. {CT-GAN}: Malicious tampering of 3d medical imagery using deep learning. In *28th USENIX Security Symposium (USENIX Security 19)*, 461–478.

Rodriguez, J.; Huang, K.; and Xu, M. 2024. Multi-Task Breast Ultrasound Image Classification and Segmentation Using Swin Transformer and VMamba Models. In *2024 7th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, 858–863. IEEE.

Seoni, S.; Jahmunah, V.; Salvi, M.; Barua, P. D.; Molinari, F.; and Acharya, U. R. 2023. Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). *Computers in Biology and Medicine*, 165: 107441.

Siegel, R. L.; Kratzer, T. B.; Giaquinto, A. N.; Sung, H.; and Jemal, A. 2025. Cancer statistics, 2025. *Ca*, 75(1): 10.

Sipola, T.; and Kokkonen, T. 2021. One-pixel attacks against medical imaging: A conceptual framework. In *World Conference on Information Systems and Technologies*, 197–203. Springer.

Su, J.; Vargas, D. V.; and Sakurai, K. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5): 828–841.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Xu, M.; Huang, J.; Huang, K.; and Liu, F. 2024. Incorporating tumor edge information for fine-grained bi-rads classification of breast ultrasound images. *IEEE Access*, 12: 38732–38744.