

Systematic Evaluation of Data Augmentation Strategies for Ordinal BI-RADS Classification in Breast Ultrasound: A Benchmark on BUS-BRA

Hadeel Saad Alghamdi¹ , Mufti Mahmud^{1,2,3} , Abdullah Faisal Al-Battal^{3,4} , and Ammar Alsheghri^{3,5} 

¹ SDAIA-KFUPM Joint Research Center for AI, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

² Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

³ IRC for Biosystems and Machines, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

⁴ Department of Electrical Engineering, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

⁵ Department of Mechanical Engineering, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

{hadeel.ghamdi, mufti.mahmud, albattal, ammar.sheghri}@kfupm.edu.sa

Abstract. Correct classification of breast ultrasound lesions into Breast Imaging Reporting and Data System (BI-RADS) categories 2–5 is clinically important yet technically challenging, particularly for the diagnostically ambiguous BI-RADS 3 category and the under-represented BI-RADS 5 category. Although data augmentation is widely used to improve deep learning models in medical imaging, its impact on full-range ordinal BI-RADS classification has not been systematically benchmarked on a public dataset. In this work, we compare six augmentation strategies—baseline, geometric, intensity, ultrasound-specific, modern, and combined—using EfficientNet-B3 and ResNet-50 trained with ordinal Earth Mover Distance (EMD) loss on the BUS-BRA dataset containing 1,875 ultrasound images from 1,064 patients. Experiments were evaluated using patient-grouped five-fold cross-validation to prevent patient-level information leakage and assessed using accuracy, macro-F1, mean absolute error (MAE), and category-specific F1 scores. For EfficientNet-B3, combined augmentation achieved the strongest overall performance (accuracy 0.602, macro-F1 0.587, MAE 0.451), whereas modern augmentation produced the highest F1 score for the minority BI-RADS 5 category. Cross-backbone evaluation further showed that augmentation effectiveness depends on network architecture, with EfficientNet-B3 benefiting consistently from all augmentation strategies, while ultrasound-specific and combined augmentation exhibited greater variability for ResNet-50. These results show that augmentation strategy should be selected according to both the target architecture and the clinically prioritised BI-RADS category rather than overall performance alone.

Keywords: Breast ultrasound · BI-RADS classification · Data augmentation · Ordinal classification · Deep learning · EfficientNet · BUS-BRA

1 Introduction

Breast cancer is the most commonly diagnosed cancer among women worldwide, and early detection continues the most effective strategy for improving survival outcomes. Breast Ultrasound (BUS) is a widely used, non-ionising, low-cost imaging modality that plays a central role in breast lesion characterisation, particularly for women with dense breast tissue or as a complement to mammography. To standardise the interpretation of breast imaging findings and guide clinical management, the American College of Radiology introduced the Breast Imaging Reporting and Data System (BI-RADS), which assigns lesions to categories reflecting increasing suspicion of malignancy: BI-RADS 2 (benign), BI-RADS 3 (probably benign, short-interval follow-up recommended), BI-RADS 4 (suspicious, biopsy recommended), and BI-RADS 5 (highly suggestive of malignancy, biopsy required). Accurate assignment of BI-RADS categories directly determines patient management, making inter-observer variability a significant clinical problem; inconsistencies contribute to both unnecessary biopsies and missed or delayed detection of malignancy.

Recent advances in deep learning have substantially improved automated breast ultrasound analysis, permitting more accurate lesion characterisation across a wide range of computer aided diagnosis tasks. However, the majority of published work simplifies the task to binary benign-vs-malignant prediction, discarding the ordinal structure that gives BI-RADS its clinical meaning. Chang et al. [3] demonstrated that predicting the full BI-RADS 2–5 range is feasible using CNN, SVM, and Random Forest classifiers on a private 151-image dataset, achieving up to 85.42% accuracy. Su et al. [16] applied ensemble DL on a larger private dataset (8,026 ROIs), achieving 81.8% accuracy but noting persistently weak performance on BI-RADS 3, a pattern that recurs across the literature. Neither study systematically examined the role of data augmentation on BI-RADS-stratified performance. Data augmentation is widely used to improve generalisation on small medical imaging datasets, yet its effectiveness varies considerably across tasks and modalities. Tupper and Gagné [18] conducted a systematic augmentation benchmark across 14 ultrasound tasks from 10 datasets spanning 11 body regions, finding that no single strategy consistently outperforms others. Critically, their benchmark does not include BUS-BRA, does not evaluate the full ordinal BI-RADS 2–5 classification, and does not report per-category metrics, leaving open the question of which strategies are most effective for this specific task.

More broadly, trustworthy medical AI increasingly requires not only strong predictive performance but also explainability, privacy, personalisation, and responsible deployment. Mahmud and collaborators have addressed these issues through explainable and privacy-preserving personalised healthcare AI [15], broader frameworks for preventive, explainable, and inclusive healthcare intelligence [13],

and human-centred privacy-preserving clinical AI [14]. These principles are particularly relevant to ultrasound decision-support systems, where clinically meaningful interpretation and reliable deployment are essential.

In this paper, we address this gap with a controlled augmentation benchmark on BUS-BRA [7], a recent publicly available multi-institutional breast ultrasound dataset providing expert BI-RADS 2–5 annotations and segmentation masks for 1,875 images. All experiments are evaluated using patient-grouped cross-validation to prevent patient-level information leakage. Our contributions are:

1. A systematic augmentation benchmark for full-range ordinal BI-RADS 2–5 classification on BUS-BRA, comparing six augmentation strategies on a publicly available, reproducible dataset. We do not claim state-of-the-art performance; rather, we provide a controlled, verifiable baseline that prior work on private datasets cannot offer.
2. An ordinal-aware training setup using Earth Mover Distance loss and class-balanced weighted sampling, with evaluation using per-category F1 specifically targeting the weak BI-RADS 3 and minority BI-RADS 5 categories.
3. Evidence that augmentation effectiveness is both category- and architecture-dependent. Combined augmentation achieved the strongest overall performance for EfficientNet-B3, whereas different augmentation strategies were optimal for individual BI-RADS categories. Evaluation using ResNet-50 further demonstrated that augmentation rankings are not necessarily transferable across backbone architectures.

2 Related Work

2.1 BI-RADS Classification with Deep Learning

Chang et al. [3] were among the first to treat BI-RADS classification as a multi-class problem rather than a binary one. They extracted 145 features covering shape, margin, echo pattern, and posterior acoustics, then compared SVM, Random Forest, and CNN on 151 private cases, with CNN reaching 85.42% across categories 2–5. Su et al. [16] later combined VGG-16, ResNet-50, and DenseNet-121 into an ensemble that achieved 81.8% on a larger private collection, though category 3 remained the weakest point — a recurring gap across the literature. Carrilero-Mardones et al. [2] took a different angle, predicting BI-RADS descriptors as an intermediate step before assigning the final risk category, reaching a Cohen’s kappa of 0.64 on proprietary data. More recently, Li et al. [10] stepped back from the full scale and focused only on category 4 using automated 3D breast volume scans, highlighting a broader tendency to narrow the problem rather than model the full ordinal range.

2.2 Explainability in Breast Ultrasound CAD

Guldogan et al. [6] combined MobileNetV2 and DenseNet-121 and applied Grad-CAM, saliency maps, and SHAP to expose model decisions to clinicians, although

their task was limited to binary malignancy classification. Explainability is particularly important in high-stakes medical AI, where black-box predictions may limit clinical trust and validation [8]. Healthcare-specific reviews likewise highlight the value of explainable modelling for understanding diagnostic predictions [19]. Our Grad-CAM analysis therefore evaluates whether the BI-RADS classifier focuses on clinically plausible lesion regions.

2.3 Data Augmentation in Medical Ultrasound

Tupper and Gagné [18] benchmarked geometric and intensity transforms across 14 tasks from 10 datasets covering 11 anatomical regions, finding augmentation effectiveness to be highly task- and dataset-dependent. Their benchmark does not include BUS-BRA, does not evaluate BI-RADS ordinal classification, and does not report per-category metrics. Our work directly extends their framing to the underexplored task of ordinal BI-RADS classification using a publicly available BUS dataset.

3 Methods

3.1 Dataset

We use BUS-BRA [7]: 1,875 BUS images from 1,064 patients acquired across four scanner models (GE Logiq 7, GE Voluson, Philips iU22, Samsung). Each image has an expert BI-RADS label confirmed by biopsy and a paired segmentation mask. The dataset contains 562 BI-RADS 2, 463 BI-RADS 3, 693 BI-RADS 4, and 157 BI-RADS 5 images, corresponding to BI-RADS 2 ($n=562$, 30.0%), BI-RADS 3 ($n=463$, 24.7%), BI-RADS 4 ($n=693$, 37.0%), and BI-RADS 5 ($n=157$, 8.4%). To obtain a clinically realistic estimate of model generalization, we adopted patient-grouped five-fold cross-validation using GroupKFold. Patient identifiers were used as grouping variables so that every image belonging to the same patient, including bilateral (left and right breast) examinations, was assigned to a single fold. Consequently, no patient contributed images to both the training and validation sets, eliminating patient-level information leakage during evaluation while preserving a similar distribution of BI-RADS categories across folds.

3.2 Augmentation Strategies

Six augmentation strategies were evaluated using the Albumentations library [1]. All images were resized to 224×224 pixels and normalised using ImageNet statistics before training. The evaluated augmentation policies are summarised in Table 2. Each strategy was predefined before experimentation and evaluated independently under an identical training protocol. No augmentation policy was selected, modified, or tuned using validation or test-fold performance. The six conditions are illustrated in Fig. 1 and listed in Table 2.

Table 1: Patient and image counts per fold after patient-grouped stratified cross-validation (GroupKFold, group = patient ID). No patient’s images appear in both training and validation partitions in any fold.

Fold	Images	Patients	BI-RADS 2	BI-RADS 3	BI-RADS 4	BI-RADS 5
1	375	212	111	98	137	29
2	375	213	108	92	145	30
3	375	213	110	93	138	34
4	375	213	118	90	140	27
5	375	213	115	90	133	37
Total	1875	1064	562	463	693	157

Augmentation strategies applied to a BUS-BRA image (BI-RADS 4)

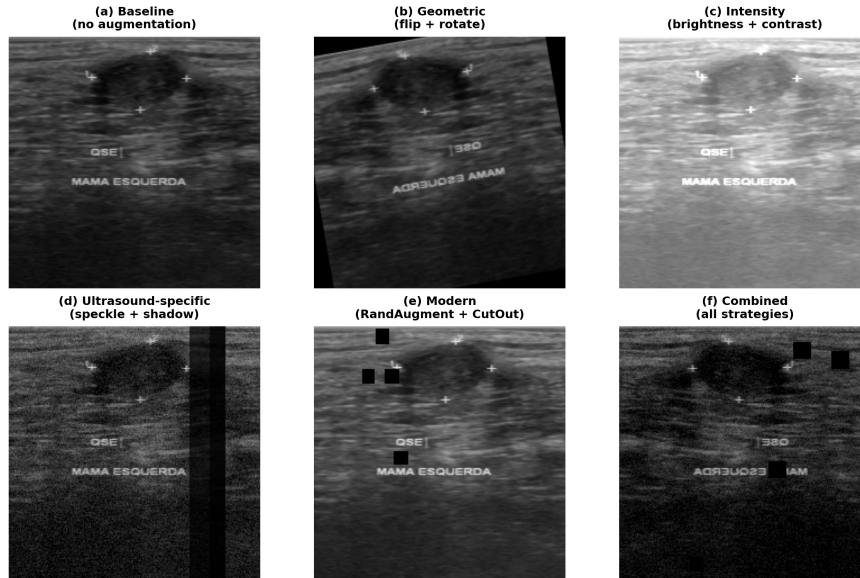


Fig. 1: Augmentation strategies applied to a BUS-BRA image (BI-RADS 4). (a) Baseline (no augmentation), (b) Geometric (flip + rotation), (c) Intensity (brightness + contrast), (d) Ultrasound-specific (speckle + shadow), (e) Modern (RandAugment + CutOut), (f) Combined (all strategies).

3.3 Model Architecture and Training

Two ImageNet-pretrained convolutional neural networks were evaluated: EfficientNet-B3 [17] and ResNet-50 [9]. For both architectures, the original classification layer was replaced with a dropout layer ($p = 0.3$) followed by a fully connected layer producing four outputs corresponding to BI-RADS categories 2–5. All experiments were performed using identical optimisation settings to ensure that augmentation strategy was the only experimental variable. Models were trained using the AdamW optimiser with an initial learning rate of 10^{-4} , weight decay of 10^{-4} , cosine annealing learning-rate scheduling, a batch size

Table 2: The six augmentation strategies evaluated. All training images were resized to 224×224 pixels and normalised using ImageNet statistics prior to augmentation.

Strategy	Transforms applied	Purpose
Baseline	Resize + normalise only	Lower-bound reference
Geometric	H-flip ($p=0.5$), V-flip ($p=0.2$), rotation $\leq 15^\circ$ ($p=0.5$), shift-scale-rotate ($p=0.4$)	Spatial diversity
Intensity	Brightness-contrast $\pm 20\%$ ($p=0.5$), gamma $\gamma \in [0.8, 1.2]$ ($p=0.4$), CLAHE ($p=0.3$)	Scanner gain variation
Ultrasound	Gaussian noise ($p=0.5$), Gaussian blur ($p=0.3$), random shadow ($p=0.3$), coarse dropout ($p=0.2$)	Clinical artefacts
Modern	Randomly selected geometric or photometric transformations [4]; CutOut regularisation [5] ($p=0.3$)	RandAugment-style
Combined	Representative geometric, intensity and ultrasound-specific transformations.	Comprehensive

of 32, and a maximum of 100 training epochs. To alleviate class imbalance, a `WeightedRandomSampler` was employed so that minority categories were sampled more frequently during training. Model selection was based exclusively on the validation fold within each cross-validation split. The model achieving the highest validation macro-F1 was retained for that fold. Early stopping with a patience of 20 epochs was applied only after a minimum training period to avoid premature termination before convergence of the minority BI-RADS 5 category.

3.4 Loss Function and Evaluation

Because BI-RADS categories are ordinal, we employ the Earth Mover Distance (EMD) loss [11], which measures the mean absolute difference between the cumulative probability distributions of the predicted and target labels, assigning larger penalties to errors that are farther from the ground truth. Performance was evaluated using accuracy, macro-F1, mean absolute error (MAE), and category-specific F1 scores for BI-RADS 2–5, reported as mean $\pm$ standard deviation across the five patient-grouped cross-validation folds. F1(BI-RADS 3) was selected as the primary category-specific metric because of its clinical importance and consistent difficulty in previous studies. Within each fold, the checkpoint with the highest validation macro-F1 was retained, while all augmentation strategies were predefined and compared only after the complete cross-validation procedure.

4 Results

4.1 Overall Performance

Table 3 summarises the performance of the six augmentation strategies using EfficientNet-B3 under patient-grouped five-fold cross-validation. All augmenta-

tion policies improved the overall classification performance compared with the baseline model, although the magnitude of improvement varied across strategies. Combined augmentation achieved the strongest overall performance, obtaining the highest accuracy (0.602), the highest macro-F1 (0.587), and the lowest MAE (0.451). Modern augmentation ranked second in macro-F1 and produced the best performance for the minority BI-RADS 5 category (see Section 4.3), while Geometric and Intensity augmentation provided comparable overall performance. These results indicate that the combination of complementary augmentation families is beneficial for the overall classification of ordinal BI-RADS, although individual augmentation strategies remain preferable for specific diagnostic categories.

Table 3: Performance of the six augmentation strategies using EfficientNet-B3 under patient-grouped five-fold cross-validation. Values are reported as mean $\pm$ standard deviation across five folds. The best result in each column is shown in **bold**.

Strategy	Acc.	Macro-F1	MAE	F1 cat2	F1 cat3*	F1 cat4	F1 cat5
Baseline	0.570 $\pm$.011	0.554 $\pm$.014	0.508 $\pm$.007	0.651	0.488	0.587	0.489
Geometric	0.586 $\pm$.024	0.571 $\pm$.021	0.477 $\pm$.024	0.649	0.520	0.610	0.507
Intensity	0.594 $\pm$.014	0.570 $\pm$.018	0.474 $\pm$.015	0.676	0.505	0.619	0.480
Ultrasound	0.597 $\pm$.021	0.570 $\pm$.023	0.466 $\pm$.019	0.666	0.524	0.624	0.466
Modern	0.586 $\pm$.027	0.576 $\pm$.024	0.479 $\pm$.024	0.663	0.498	0.604	0.539
Combined	0.602 $\pm$.014	0.587 $\pm$.025	0.451 $\pm$.007	0.665	0.529	0.627	0.529

Table 4: Fold-wise performance of the best-performing augmentation strategy (Combined) using EfficientNet-B3.

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy	0.600	0.579	0.608	0.616	0.605
Macro-F1	0.592	0.544	0.597	0.605	0.598
MAE	0.448	0.461	0.451	0.443	0.453
F1 cat2	0.673	0.646	0.647	0.654	0.703
F1 cat3	0.566	0.525	0.556	0.502	0.495
F1 cat4	0.586	0.625	0.636	0.676	0.612
F1 cat5	0.542	0.381	0.551	0.588	0.583

4.2 Comparison with Prior Work

Table 5 places our benchmark in the context of previous full-range BI-RADS classification studies. Because the studies differ in datasets, evaluation protocols, and reported metrics, the results should be interpreted qualitatively rather than as direct numerical comparisons. A more detailed discussion is provided in Section 5.

Table 5: Comparison with representative full-range BI-RADS classification studies. Differences in datasets and evaluation protocols should be considered when interpreting the reported results.

Study	Dataset	Images	Public?	Eval.	Metric	Result
Chang et al. [3]	Private	151	No	Train/test [‡]	Accuracy	85.4%
Su et al. [16]	Private	8,026 [†]	No	Train/test [‡]	Accuracy	81.8%
Ours (Combined)*	BUS-BRA	1,875	Yes	Patient-grouped 5-fold	Macro-F1	58.7%
Ours (Combined)*	BUS-BRA	1,875	Yes	Patient-grouped 5-fold	Accuracy	60.2%

4.3 Per-Category Analysis

Category-specific F1 scores show that augmentation strategies influence individual BI-RADS categories differently, indicating that overall performance alone does not fully characterize model behaviour.

- **BI-RADS 2 (benign):** Intensity augmentation achieved the highest F1 score (0.676), closely followed by Ultrasound (0.666) and Combined (0.665).
- **BI-RADS 3 (probably benign):** behavior Combined augmentation produced the highest F1 score (0.529), with Ultrasound (0.524) and Geometric (0.520) showing comparable performance.
- **BI-RADS 4 (suspicious):** Combined augmentation achieved the highest F1 score (0.627), followed by Ultrasound (0.624) and Intensity (0.619).
- **BI-RADS 5 (highly malignant):** Modern augmentation obtained the highest F1 score (0.539), narrowly ahead of Combined (0.529) and Geometric (0.507). In contrast, the baseline model achieved an F1 score of only 0.489, highlighting the importance of augmentation to improve recognition of the minority class.

Overall, these results demonstrate that no single augmentation strategy optimizes performance across all BI-RADS categories. Consequently, the choice of augmentation should be guided by the clinical objective rather than by a single aggregate performance metric.

4.4 Grad-CAM Explainability

Figure 2 presents representative Grad-CAM visualizations generated by the Combined EfficientNet-B3 model using one correctly classified validation image from each BI-RADS category. Explainable medical AI should support clinically interpretable reasoning rather than relying solely on predictive performance [8,19]. Across all categories, the network consistently focuses on the lesion rather than surrounding background tissue. For BI-RADS 2 and 3, attention is distributed across the lesion and its immediate boundary, whereas BI-RADS 4 and 5 exhibit more concentrated activation around the lesion margin and posterior acoustic region. These observations suggest that the model relies on clinically relevant lesion characteristics when making its predictions.

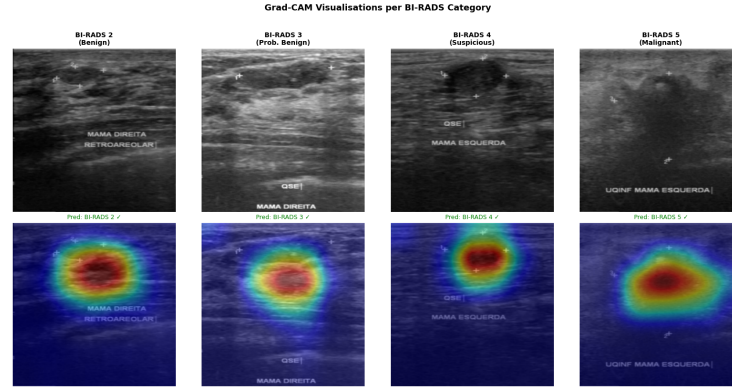


Fig. 2: Grad-CAM visualizations for one representative image per BI-RADS category using the Combined augmentation model. Top row: original ultrasound images. Bottom row: gradient-weighted activation maps overlaid on the image. Warmer colours indicate regions contributing most strongly to the model prediction.

4.5 Performance on Clinically Important Categories

Figure 5 compares augmentation performance for the two clinically important BI-RADS categories. For the minority BI-RADS 5 category, Modern augmentation achieved the highest F1 score (0.539), closely followed by Combined (0.529) and Geometric (0.507). Although the baseline already achieved reasonable BI-RADS 5 performance ($F1 = 0.489$), appropriate augmentation further improved minority-class recognition. For BI-RADS 3, Combined augmentation produced the highest F1 score (0.529), followed closely by Ultrasound (0.524). These findings demonstrate that augmentation strategies influence individual BI-RADS categories differently, emphasizing the importance of category-specific evaluation rather than relying solely on overall accuracy or macro-F1.

4.6 Statistical Analysis

The statistical significance of the observed augmentation effects was evaluated using the Wilcoxon signed-rank test applied to paired macro-F1 scores from the five patient-grouped cross-validation folds. None of the pairwise comparisons reached the conventional significance threshold ($\alpha = 0.05$). This outcome is expected given the limited statistical power associated with only five paired observations, for which the minimum attainable p -value is 0.0625 [20]. Nevertheless, all augmentation strategies improved macro-F1 relative to the baseline, with Combined augmentation achieving the largest overall improvement (macro-F1 = 0.587). Similar trends were observed for the minority BI-RADS 5 category, where Modern ($F1 = 0.539$) and Combined ($F1 = 0.529$) produced the strongest performance. These results indicate consistent practical improvements, although larger datasets or repeated cross-validation are required to establish statistical significance with greater confidence.

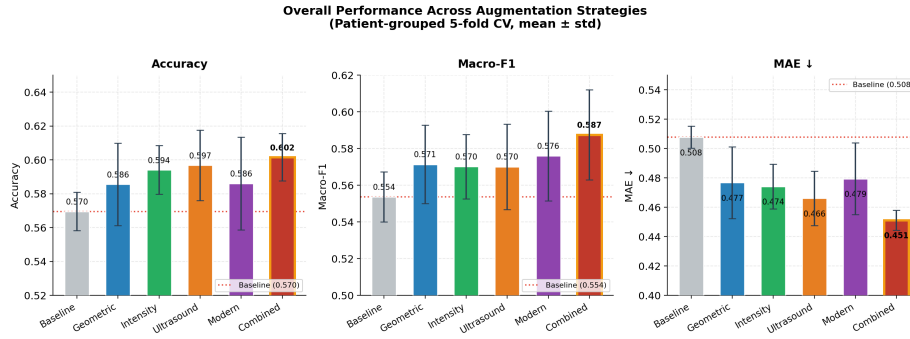


Fig. 3: Overall performance of the six augmentation strategies using EfficientNet-B3 under patient-grouped five-fold cross-validation. Bars represent the mean performance across the five folds, with error bars indicating one standard deviation. The dashed red line denotes the baseline model without augmentation.

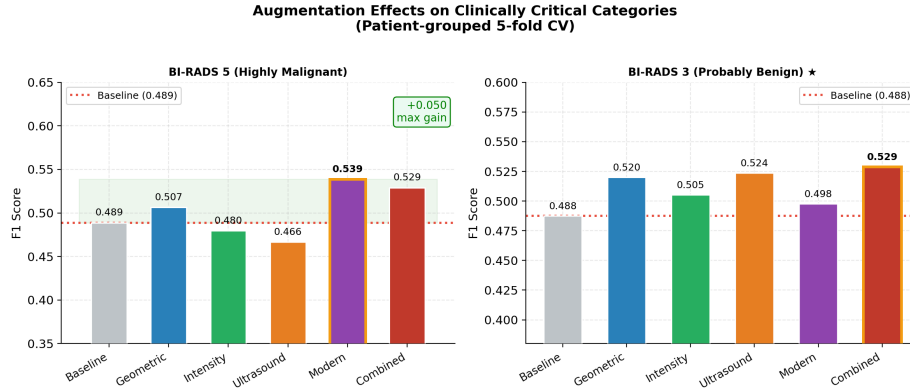


Fig. 4: Category-specific F1 scores obtained using EfficientNet-B3. The heatmap (left) and grouped bar chart (right) illustrate that different augmentation strategies favour different BI-RADS categories, demonstrating that no single augmentation policy is optimal across all classes.

4.7 Architecture Robustness

To investigate whether augmentation performance generalizes across network architectures, the complete benchmark was repeated using ResNet-50 [9]. Unlike EfficientNet-B3, where Combined augmentation achieved the strongest overall performance (macro-F1 = 0.587), Modern augmentation produced the highest macro-F1 for ResNet-50 (0.578), with Geometric ranking second. The most pronounced architectural differences were observed for Ultrasound and Combined augmentation. While both strategies produced competitive results with EfficientNet-B3, they exhibited substantially lower performance for ResNet-50, particularly for the minority BI-RADS 5 category. In contrast, Baseline, Geometric, and Intensity augmentation showed more consistent behavior across

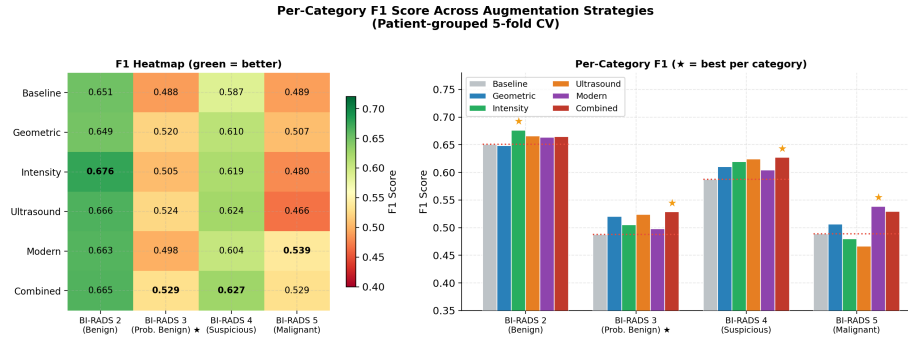


Fig. 5: Comparison of augmentation performance for the two clinically important categories, BI-RADS 5 (left) and BI-RADS 3 (right). Modern augmentation achieved the highest BI-RADS 5 performance, whereas Combined augmentation produced the highest F1 score for BI-RADS 3 under the patient-grouped evaluation protocol.

both backbones. These findings indicate that augmentation effectiveness is influenced not only by the augmentation policy itself but also by the feature-learning characteristics of the underlying CNN architecture. Consequently, augmentation strategies should be validated jointly with the selected backbone rather than assumed to transfer unchanged across different models.

Table 6: Performance comparison of EfficientNet-B3 and ResNet-50 across all augmentation strategies under patient-grouped five-fold cross-validation

Strategy	Backbone	Acc.	Macro-F1	MAE	cat2	cat3	cat4	cat5
Baseline	EffNet-B3	0.570	0.554	0.508	0.651	0.488	0.587	0.489
Baseline	ResNet-50	0.575	0.542	0.510	0.647	0.456	0.625	0.439
Geometric	EffNet-B3	0.586	0.571	0.477	0.649	0.520	0.610	0.507
Geometric	ResNet-50	0.591	0.577	0.463	0.675	0.513	0.609	0.510
Intensity	EffNet-B3	0.594	0.570	0.474	0.676	0.505	0.619	0.480
Intensity	ResNet-50	0.599	0.558	0.463	0.683	0.511	0.631	0.405
Ultrasound	EffNet-B3	0.597	0.570	0.466	0.666	0.524	0.624	0.467
Ultrasound	ResNet-50	0.578	0.470	0.479	0.524	0.499	0.681	0.175
Modern	EffNet-B3	0.586	0.576	0.480	0.663	0.498	0.604	0.539
Modern	ResNet-50	0.605	0.578	0.454	0.672	0.527	0.640	0.474
Combined	EffNet-B3	0.602	0.587	0.451	0.665	0.529	0.627	0.529
Combined	ResNet-50	0.567	0.469	0.479	0.545	0.511	0.635	0.186

5 Discussion

Augmentation remains beneficial. Although the patient-grouped evaluation protocol produced a stronger baseline than our initial experiments (macro-F1 = 0.554), all evaluated augmentation strategies further improved overall performance. Combined augmentation achieved the highest macro-F1 (0.587), confirm-

ing that carefully designed augmentation remains beneficial for ordinal BI-RADS classification on BUS-BRA.

No single strategy wins everywhere. Combined augmentation achieved the best overall performance, but the category-specific analysis reveals a more nuanced picture. Intensity augmentation performed best for BI-RADS 2, Combined for BI-RADS 3 and BI-RADS 4, and Modern for BI-RADS 5. Consequently, the most appropriate augmentation strategy depends on the clinical objective rather than a single overall performance metric. For applications prioritizing the detection of highly malignant lesions, Modern augmentation provides the strongest BI-RADS 5 performance, whereas Combined augmentation is preferable for improving classification of the diagnostically challenging BI-RADS 3 category.

More augmentation is not always better. Although combined augmentation achieved the strongest overall performance for EfficientNet-B3, the same behavior was not observed for ResNet-50, where Modern and Geometric augmentation performed better. This suggests that the effectiveness of augmentation depends not only on the augmentation policy itself but also on the underlying network architecture.

How our numbers compare to prior work. Our best accuracy of 0.602 is lower than the figures reported by Chang et al. [3] (85.4%) and Su et al. [16] (81.8%), and we think three differences explain this gap. Their datasets were either very small, single-hospital, or used pre-cropped lesion patches rather than full ultrasound frames. BUS-BRA is a harder setting: full frames from four different scanner models. They also used simple train-test splits on private data, while we use patient-grouped five-fold cross-validation on a public dataset, which typically produces lower but more honest numbers. Finally, we report macro-F1 rather than accuracy, which penalizes poor performance on minority classes like BI-RADS 5 more severely.

Clinical AI perspective. The findings should also be interpreted within the broader movement towards human-centred and inclusive healthcare AI. Recent work has emphasised privacy-preserving personalised systems [15], human-centred multi-agent healthcare intelligence [14], and explainable multimodal digital-human modelling for accessibility and inclusion [12]. Accordingly, future BI-RADS decision-support systems should combine performance improvements with transparency, robustness, and clinically appropriate human oversight.

Limitations. We tested two backbones (EfficientNet-B3 and ResNet-50) which exhibited different augmentation rankings; other loss functions or architectures may still yield different outcomes. The dataset contains only 1,875 images, which limits the statistical power of the Wilcoxon analysis despite consistent improvements across the cross-validation folds. Furthermore, the differing behavior of EfficientNet-B3 and ResNet-50 indicates that augmentation strategies should be validated across multiple architectures.

Future work. Future work should evaluate these augmentation strategies on additional multi-centre breast ultrasound datasets and scanner types, and extend the benchmark to transformer-based and multimodal architectures. Beyond predictive performance, future systems should incorporate explainability, uncertainty, privacy, and inclusive human-centred design, consistent with emerging trustworthy healthcare AI frameworks [13,12].

6 Conclusion

We presented a systematic benchmark of six data augmentation strategies for ordinal BI-RADS 2–5 classification on the publicly available BUS-BRA breast ultrasound dataset. Combined augmentation achieved the strongest overall performance for EfficientNet-B3, whereas Modern augmentation produced the highest performance for the minority BI-RADS 5 category. The category-specific analysis further demonstrated that no single augmentation strategy is optimal across all BI-RADS categories, highlighting the importance of selecting augmentation policies according to the intended clinical application. Evaluation using a second backbone showed that augmentation effectiveness is architecture-dependent, with several augmentation strategies exhibiting substantially different behavior for EfficientNet-B3 and ResNet-50. These findings suggest that augmentation policies should be selected jointly with the target network architecture rather than assumed to generalize across different models. We hope that the proposed benchmark and publicly available implementation will provide a reproducible baseline for future research in ordinal breast ultrasound classification.

Code and Data Availability

The source code used to reproduce the experiments presented in this paper is publicly available at <https://github.com/hillihadeel940-AI-Land/busbra-augmentation-benchmark>. The BUS-BRA dataset is publicly available from the original authors [7]. Additional materials and trained models are available in the project repository.

References

1. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: Fast and flexible image augmentations. *Information* **11**(2), 125 (2020)
2. Carrilero-Mardones, O., et al.: Deep learning for describing breast ultrasound images with BI-RADS terms. *Journal of Imaging Informatics in Medicine* (2024). <https://doi.org/10.1007/s10278-024-01013-0>
3. Chang, Y.W., Chen, Y.R., Ko, C.C., Lin, W.Y., Lin, K.P.: A novel computer-aided-diagnosis system for breast ultrasound images based on BI-RADS categories. *Applied Sciences* **10**(5), 1830 (2020)

4. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: RandAugment: Practical automated data augmentation with a reduced search space. In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 33, pp. 18613–18624 (2020)
5. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552* (2017)
6. Guldogan, N., et al.: Artificial intelligence in BI-RADS categorization of breast lesions on ultrasound: Can we omit excessive follow-ups and biopsies? *Academic Radiology* **31**(6), 2194–2202 (2024)
7. Gómez-Flores, W., Gregorio-Calas, M.J., Coelho de Albuquerque Pereira, W.: Busbra: A breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics* **51**(4), 3110–3123 (2024)
8. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A.: Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation* **16**(1), 45–74 (2024)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 770–778 (2016)
10. Li, Y., et al.: Multiview deep learning networks based on automated breast volume scanner images for identifying breast cancer in BI-RADS 4. *Frontiers in Oncology* **14**, 1399296 (2024)
11. Litvinov, A., Ushakov, E., Senotrusova, S., Lukianov, K., Markin, Y., Mikhailova, L., Karpulevich, E.: Clinically aware learning: Ordinal loss improves medical image classifiers. *Journal of Clinical Medicine* **15**(1), 365 (2026)
12. Mahmud, M., Baslyman, M., et al.: Inclusive ai for digital human modelling: A perspective on explainable multimodal digital twins for accessibility and inclusion. In: *International Conference on Human-Computer Interaction*. pp. 1–16. Springer (2026)
13. Mahmud, M., Brown, D.J., Shen, Y., et al.: Preventive healthcare through privacy-preserving, explainable and inclusive artificial intelligence. In: *Lecture Notes in Computer Science*. vol. 16340 LNCS, pp. 359–374 (2026)
14. Mahmud, M., Shaffi, N., Kaiser, M.S., et al.: Privacy-preserving multi-agent systems for human-centred inclusive healthcare ai. *AI Magazine* **47**(3), e70087 (2026)
15. Mahmud, M., et al.: Towards explainable and privacy-preserving artificial intelligence for personalisation in autism spectrum disorder. *Lecture Notes in Computer Science* **13309 LNCS**, 356–370 (2022)
16. Su, C., et al.: A practical CAD system for breast ultrasound classifying lesions into ACR BI-RADS assessment categories. *Journal of Medical and Biological Engineering* (2024)
17. Tan, M., Le, Q.V.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. pp. 6105–6114 (2019)
18. Tupper, A., Gagné, C.: Revisiting data augmentation for ultrasound images. *arXiv preprint arXiv:2501.13193* (2025)
19. Viswan, V., Shaffi, N., Mahmud, M., Subramanian, K., Hajamohideen, F.: Explainable artificial intelligence in alzheimer’s disease classification: A systematic review. *Cognitive Computation* **16**(1), 1–44 (2024)
20. Wilcoxon, F.: Individual comparisons by ranking methods. *Biometrics Bulletin* **1**(6), 80–83 (1945)