

Integrating Median Filtering with Deep Pattern Recognition Models for Improved Breast Ultrasound Cancer Detection

Zeeshan Ahmad^{1*}, and Yasir Aziz¹

¹Department of Computer Engineering, Bahauddin Zakariya University, Multan, Pakistan.

*Corresponding Author: Zeeshan Ahmad. Email: engr_zeeshan@hotmail.co.uk

Received: May 01, 2026 Accepted: September 10, 2026

Abstract: Even though ultrasound of the breast is an easy to access imaging modality and is radiation free, speckle-like intensity fluctuations, low contrast, variability of acquisition and limited model transparency are the challenges in automated classification of ultrasound lesions. In this paper, we introduce a controlled framework of preprocessing-classification-explainability that compares the aforementioned three image types (raw, 3×3 median filtered, median filtered + contrast-limited adaptive histogram equalization (CLAHE)) using the same leakage-safe data split, convolutional neural network (CNN), training budget and evaluation protocol. To minimize the information leakage, the downloaded Kaggle archive of benign and malignant images was pre-augmented with 9016 images; and then divided into 70/15/15 train/validation/test sets based on augmentation-family grouping. The 96.97% accuracy, 98.95% sensitivity, 95.04% specificity, 95.10% precision, 96.99% F1-score, 0.99623 ROC-AUC and 0.99608 PR-AUC of the model that was median-filtered were the highest. Compared to raw standardized baseline, median filtering resulted in an increase in accuracy and sensitivity by 3.55 and 6.15 percentage points, respectively, and in a decrease of false negative cases from 48 to 7. Median restoration also gave the best results in terms of PSNR and SSIM for controlled synthetic-speckle testing, which were 27.9313 dB PSNR and 0.8214 SSIM, respectively. There was a significant reduction in paired errors after raw versus median filtering (adjusted $p = 0.000008$) by Exact McNemar analysis. Grad-CAM, LIME and SHAP also revealed class-specific areas and failure modes but only showed low spatial overlap, which indicated that these are complementary rather than interchangeable. The results indicate that a very basic median filter which preserves the edges can enhance the predictive reliability and model auditing when the effects of the pre-processing are separated in an experimental setup. There was only one fixed 70/15/15 split in which the family-aware group was divided, so there is no uncertainty estimation based on repeated cross-validation, but rather on bootstrap resampling on the held-out test cohort.

Keywords: Breast Ultrasound; Median Filtering; Convolutional Neural Network; Speckle Noise; CLAHE; Explainable Artificial Intelligence; Grad-CAM; LIME; SHAP

1. Introduction

Breast ultrasound is an adjunct to mammography because of its portability, relative low cost, ability to be repeated and lack of ionizing radiation. It is particularly useful for detecting abnormalities in the breast when the breast has a lot of fibroglandular tissue that makes the abnormality on the mammogram hard to spot. However, ultrasound is not a standard imaging source per se. The visual impression of a lesion will vary depending on the parameters of a probe such as pressure, scanning angle, gain, the depth of focus, patient anatomy and the properties of the device. Besides, coherent interference creates a speckle pattern with its granular structure which may obscure the small margins or internal structure. These are the reasons why it is challenging to distinguish between malignant and benign masses, for both clinicians and

automated systems. The task can thus not only be learning discriminative lesion-sets but also controlling the variation of the input set on which the discriminative lesion sets is learned [1] [2].

There are two general approaches to solve this problem which are currently being used by the existing computer aided diagnosis (CAD) systems. The first involve classical image processing techniques that are applied to reduce noise, enhance contrast or select a ROI prior to classification. These methods are fast, easy to interpret but can be aggressive, destroying diagnostically important edges, or can enhance contrast which can obscure background texture. The second approach is deep learning that uses the CNN, transfer learning models, vision transformers, and ensembles to directly learn hierarchical representations from image data. Recent studies have suggested that classification can be enhanced by using multiple time-domain feature maps [2]–[5] and knowledge distillation [10] as well as by combining images with multi-task learning. A lot of research, however, is done by having both augmented variants and duplicates within training and testing sets, which can overestimate performance and many of these studies vary the preprocessing as well as the model architecture, making it challenging to determine the contribution of each.

There are several recent methods which have tried to combine denoising and classification. Alotaibi et al. used speckle filtering, region highlighting, RGB fusion and transfer learning on a variety of ultrasound databases [4]. Wang et al. directly attached a speckle-suppression network to an ultrasound classifier [7] while Hayum et al. had a pipeline of a fuzzy CNN network [6]. These studies show that it is worthwhile to optimize the image quality along with the classification simultaneously. However, complex pipelines can result in segmentation dependence or significant amounts of computation or even additional trainable elements which make it hard to determine if a simpler preprocessing step would have a similar effect. Hence it is necessary to conduct a controlled experiment to quantify the effect of preprocessing itself, while fixing the model, data membership, augmentation and duration of training.

In medical imaging transparency is an additional requirement. High classification scores do not indicate if the network was based on the lesion, an acquisition artifact, overlaid text or background tissue. Thus, domain-informed interpretable systems, as well as post-hoc visual methods, have become crucial in the breast ultrasound research field [9] [12] [13]. Three methods, grad-CAM [16], LIME [17] and SHAP [18] are proposed to use for prediction contribution distribution: grad-CAM based on class gradients highlights influential convolutional regions, LIME based on perturbed super pixels learns a local surrogate model, and SHAP based on Shapley value principles distributes prediction contribution. The different techniques provide explanations of different mathematical objects; one cannot assume that they agree. All three options give a more robust audit option than a single pretty heatmap.

This current study is based on an archive of ultrasound images of breasts for the purpose of breast cancer, obtained from the Kaggle website. The downloaded archive consisted of 9,016 already augmented images into two classes of benign and malignant as opposed to the original 780 images BUSI release by Al-Dhabyani et al. [1]. Rotated and sharpened variants were part of the archive – which meant that on normal random splits, there was a high risk of leakage. The methodology thus starts with the source-family audit and 70/15/15 split (leakage safe) and then three paired input branches, consisting of raw standardized, median filtered and median filtered plus CLAHE. All branches are trained with the same number of epochs, in this case 15 epochs are used for the custom 4-block CNN. Performance is then measured by validation-only threshold selection, an independent test evaluation, 2,000-resample bootstrap intervals and exact paired McNemar tests.

There are four contributions of this paper. First, it offers a family friendly Audit of a pre-augmented breast ultrasound archive, and eliminates the possibility of variant crossing between experimental subsets. Second, it does not vary any of the preprocessing factors, such as architecture, training data, augmentation, class weighting or optimization because it has a matched branch-wise design. Third, it goes beyond image-quality evidence, and presents clinically relevant test metrics, calibration measures, confidence intervals and paired significance testing. Fourth, it corrects predictions and errors with Grad-CAM, LIME and SHAP, and quantifies the concentration of explanations, as well as the spatial agreement between the methods. The output of the chain of evidence is connected to the quality of the pre-processing, the classification behavior and finally to the interpretation of the model.

The rest of the paper is structured in the following way. In section 2, we perform a review of the most recent research on breast ultrasound classification, preprocessing and explainability, highlighting the

remaining methodological challenges. The audited data set and preprocessing branches, CNNs, mathematical formulation, evaluation, and XAI procedures are described in Section 3. Results for image quality, training, testing, statistical, explainability and comparison are reported in section 4. The paper ends in the last section with some suggestions for future clinical and technical development.

2. Literature Review

From handcrafted texture descriptors to end-to-end deep representation learning, automated breast ultrasound analysis has come a long way. Today's problem is not just one of maximizing accuracy; it's also about: curating datasets; being speckle- and acquisition-variation-resistant; error analysis for clinically meaningful errors; uncertainty; and explanation. Although the original BUSI resource allowed for open comparison, it has been found that care needs to be taken when dealing with public collections of ultrasound images as they can be biased through duplicated or augmented images [1] [10].

2.1. Deep Models for Breast Ultrasound Classification

CNN based systems continue to be effective since the local convolutional filters are capable of picking up margins, internal texture, posterior acoustic patterns and heterogeneous echogenicity. In a small data setting, Kim et al. added entropy and phase feature maps to B-mode input which provides the classifier with additional information [2]. Ali et al. used a combination of various transfer-learning models by utilizing a meta learning ensemble and reported around 90% accuracy on BUSI [3]. Alotaibi et al. applied image fusion and transfer learning with VGG19 and showed that with the preprocessing it was better to predict in public and institutional datasets [4]. However, these studies demonstrate that complementary representations can be beneficial for robustness, and they tend to include multiple simultaneous changes (from new inputs, to transfer learning, augmentation and ensemble logic).

Transformer-based and hybrid models go beyond local convolution to have a larger receptive field. Zhou and Mosadegh achieved an ensemble of vision transformers with an AuROC of 0.977 and distilled into a single model with an AuROC of 0.972 [5]. Fusion of CNN is more recent method, where local and global features are explicitly exchanged and have obtained competitive internal and external validation. Unfortunately, their representational capacity is their strength, their number of parameters, cost of training and sensitivity to the number of dimensions in the dataset may hinder deployment in resource-constrained applications. By contrast, a small custom CNN may be useful in the case where the research question is related to the preprocessing and not novelty in the architecture.

2.2. Preprocessing, Denoising, and Multi-task Learning

Noise-aware studies suggest that things could change downstream as a result of ultrasound preprocessing. Hayum et al. [6] employed bilateral filtering, contrast stretching, fuzzy clustering, feature selection and fuzzy CNN; and obtained around 95% accuracy, based on a mammography-based evaluation. Wang et al. combined the speckle suppression and classification in a single trainable framework, and highlighted that the suppression of speckle noise should maintain task-relevant features [7]. Islam et al. used an ensemble CNN with U-Net segmentation, and Grad-CAM and achieved 87.82% accuracy on one ultrasound dataset and 85.69% accuracy on another [8]. These can be effective approaches, but they entail a higher level of complexity with segmentation masks, large ensembles, or multi-stage optimization, and necessitating ablation.

Additionally, dataset curation and multi-task learning have become the main. Aumento-Maestro et al. identified that BUSI images contained some duplicated images and created an end-to-end approach for the classification and segmentation of images, which was found to achieve improvements of around 15% over single task approaches [10]. Not only is their work relevant, but it demonstrates that irregularities in the data can have an impact on performance as can architecture design. This current study uses the same general idea of auditing the datasets but serves to answer a different question: does a simple median filter help a fixed classifier in the case where related families of images are still in the same subset.

2.3. Explainability and Clinical Interpretability

Interpretability research comes in various forms from post-hoc saliency to models centered on clinically relevant prototypes. Yan et al. created an interpretable multimodal ultrasound system based on 4,320 images of 1,440 biopsy verified lesions and were able to achieve an AUC of 0.902 with confidence boosting

and decision support for the radiologists [9]. In [12] Alom et al. combined the DenseNet121, class weighting and Grad-CAM to an XAI based breast cancer framework. Ganie et al. proposed DRM-Net which is a combination of transfer-learning models, and achieved 96.71% accuracy, 97% recall and 0.99 AUC [13]. Although these studies reveal the usefulness of evidence transparency, many of them present only some of the correct cases and fail to quantify if various XAI methods highlight the same kind of regions.

The three methods – Grad-CAM, LIME and SHAP are complementary. The model specific and spatially coarse Grad-CAM is directly connected to convolutional gradients [16]. The level of super pixels gives a model-agnostic and intuitive explanation, while the level of perturbation sampling can be different as segmentation can be different [17]. SHAP is an additive-attribution basis that is principled, but relies on the background distribution and is costly for images [18]. So, a thorough XAI analysis ought to account for both the right and wrong predictions, provide report configuration information and deprecate the use of heatmaps as clinical segmentation masks.

Selected literature demonstrates good progress but has common weaknesses: non-equivalent datasets and classes, classifiable augmented data may be split randomly, external validation is rarely performed, architecture may be confounded, only accuracy is reported, and XAI is qualitative only, lacking agreement analysis. The current study attempts to overcome these problems with a leakage safe matched comparison, extensive metric reporting, paired error testing and multi-method explanation. Representative studies' strengths, findings and limitations are summarized in Table 1.

Table 1. Representative literature on breast ultrasound preprocessing, classification, and explainability

Article/Ref.	Dataset	Contribution / Method	Results / Finding	Weakness / Limitation
Kim et al. [2]	BUSI, 780 images	CNN with B-mode, entropy, and phase maps	Feature-map fusion improved most CNN variants	Small public dataset; no external test
Ali et al. [3]	BUSI	Meta-learning ensemble of pretrained CNNs	About 90% accuracy	Ensemble complexity; limited leakage discussion
Alotaibi et al. [4]	BUSI, Dataset B, KAIMRC	Speckle filtering, ROI highlighting, RGB fusion, VGG19	Preprocessing improved multi-dataset prediction	Multiple coupled preprocessing steps
Zhou and Mosadegh [5]	Curated BUSI	ViT ensemble and knowledge distillation	AuROC 0.977 ensemble; 0.972 distilled	Cross-validation; limited image-quality analysis
Hayum et al. [6]	MIAS/DDSM	Bilateral filtering, fuzzy clustering, FCNN	About 95% accuracy	Mammography rather than BUS; complex pipeline
Wang et al. [7]	Ultrasound datasets	Joint speckle suppression and classification	Noise-aware joint learning improved diagnosis	Trainable denoising increases complexity
Islam et al. [8]	Two ultrasound datasets	EDCNN, U-Net, Grad-CAM	87.82% and 85.69% accuracy	Segmentation dependence; no paired error test
Yan et al. [9]	4,320 multimodal images	Domain-knowledge prototype network	AUC 0.902; radiologist assistance benefit	Private multimodal cohort; not directly reproducible
Aumente-Maestro et al. [10]	Curated BUSI	Duplicate detection and multi-task classification/segmentation	Near 15% improvement over single-task models	Requires masks and multitask labels

Saini et al. [11]	Two public + one in-house BUS datasets	2D-VMD guided CNN with attention	98%, 93%, and 89% accuracy	Advanced decomposition; cross-dataset heterogeneity
Alom et al. [12]	BUSI and BreakHis	DenseNet121-based DNBCD with Grad-CAM	89.87% on BUSI	Three-class protocol differs from binary studies
Ganie et al. [13]	Breast ultrasound dataset	Hybrid DRM-Net transfer learning and CAM	96.71% accuracy; 97% recall	Large stacked model; 2025 single- study evidence
Present study	Kaggle archive, 9,016 images	Family-aware split; raw/median/CLAHE matched CNN; 3 XAI methods	96.97% accuracy; 98.95% sensitivity; 0.99623 AUC	Single public archive; no masks or clinical validation

3. Methodology

3.1. Dataset Description and Leakage-safe Design

The data for the experiment has been downloaded from Kaggle with the identifier ultrasound-breast-images-for-breast-cancer. It was made up of 9,016 benign and malignant image files; 4,574 benign and 4,442 malignant image files. The images were mostly 224×224 pixels with a few being 227×227 pixels, and were all decoded as three channel images even though they were grayscale ultrasound images. When downloaded, it was found that this collection was different from the original 780 image, three class BUSI collection [1] and lacked a normal class and segmentation masks. Hence, the paper only reports on binary classification. The number of benign and malignant cases were almost equal for the two classes (4,574 vs. 4,442). Patient-level independence and representativeness of the subgroups could not be confirmed based on downloaded data, as the archive provided no patient identifiers, scanner/site metadata, labels for lesion subtypes or demographic variables.

Variants were detected in the archive using the filename inspection, including some rotated and sharpened images. Images were grouped by source-family before splitting, to prevent related variants from being split into different subsets. The entire family was forced into one family group. This resulted in almost balanced classes with 6310 training images, 1353 validation images and 1353 independent test images. This distribution is given in Table 2. The augmentation families were deduced from the stems of the filenames and from the augmentation suffixes (e.g. rotated and sharpened variants) and each detected variant of a given family was given the same family identifier, while the complete family kept within one split. This results in a decreased amount of leakage at the variant level, but does not replace true patient identifiers, which come from the filename.

Table 2. Leakage-safe dataset distribution used in the experiment

Subset	Benign	Malignant	Total	Proportion
Training	3,202	3,108	6,310	69.99%
Validation	686	667	1,353	15.01%
Independent test	686	667	1,353	15.01%
Total	4,574	4,442	9,016	100.00%

3.2. Standardization and Preprocessing

Each image was resized to 224×224 pixels and normalized to $[0,1]$. If x denotes an 8-bit pixel intensity, normalization was performed according to (1).

$$x_{\text{norm}} = x / 255 \quad (1)$$

Three paired image representations were produced. The raw standardized branch applied only resizing and format conversion. The median branch converted the image to grayscale, used a 3×3 neighborhood, and replaced each central pixel by the median of neighboring intensities. For neighborhood Ω around position (i,j) , the operation is defined in (2).

$$I_{\text{med}}(i,j) = \text{median} \{ I(m,n) : (m,n) \in \Omega(i,j) \} \quad (2)$$

The third branch was using CLAHE with median filtering with clip limit 2.0 and 8×8 tile grid. CLAHE redistributes intensities in the locality while ensuring that the number of bins with high number of intensities do not contribute too much to the histogram, thus enhancing the local contrast without any unrestricted amplification. Each branch obtained as a result of the grayscale was also converted to three channels and a $224 \times 224 \times 3$ CNN input was maintained. The entire process of auditing, preprocessing, training, independent testing and explainability is summarized in Figure 1.

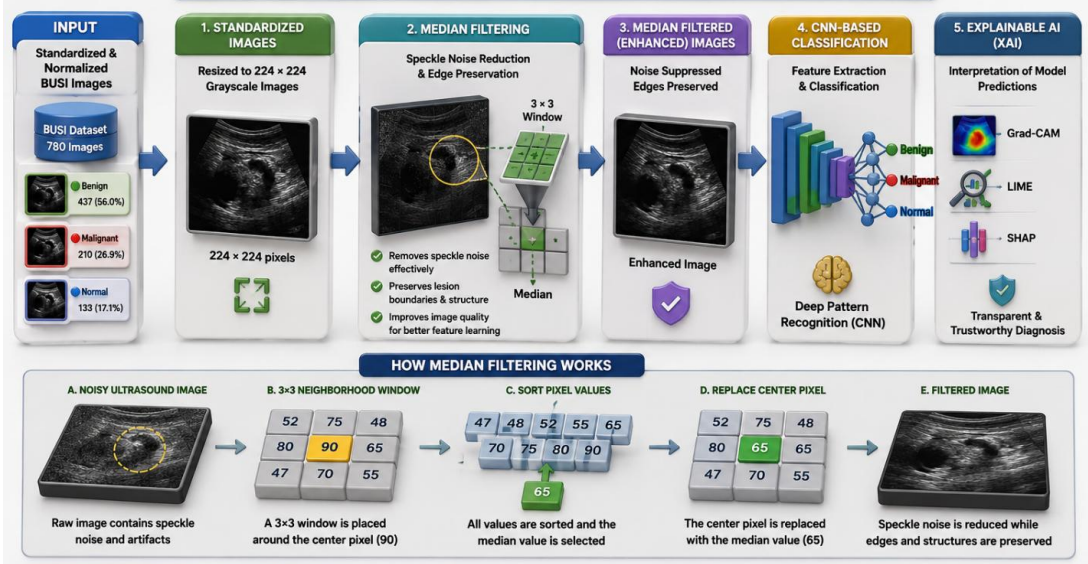


Figure 1. Complete methodology for leakage-safe preprocessing, matched CNN training, independent evaluation, and explainability

3.3. Image-quality Evaluation

Structural similarity relative to the raw standardized image, proxy contrast-to-noise ratio (CNR), entropy, RMS contrast, edge preservation, Laplacian sharpness and Equivalent Number of Looks (ENL) were used for the image level comparison of the preprocessing branches. Lesion masks were not included in the downloaded archive and CNR was automatically calculated from contrast-dominant and background-like regions and thus is reported here as proxy CNR. Standardized images also were degraded by adding synthetic multiplicative speckle to serve as a known degraded reference. PSNR, SSIM, Proxy CNR, edge preservation and ENL were computed between the output of the median and median + CLAHE and the clean image. If there was a difference between the branches, paired Wilcoxon signed-rank tests using Holm correction were used to determine whether the differences were systematic. The synthetic-speckle analysis was not a simulation of an actual ultrasound scanner but a carefully designed stress test: The clean image was used as a reference image for each of the synthetic ones, no speckle was added to the reference image, and the image was "cleared" in the median and median-plus-CLAHE branches, prior to comparison with the reference image. PSNR measures reconstruction fidelity at the pixel level, while SSIM reflects the luminance, contrast and structural similarity, with higher values signifying better restoration in this controlled perturbation but not necessarily denoising performance for all clinical acquisitions.

3.4. CNN Architecture and Training

To separate out the effects of the preprocessing, a custom four-block CNN was used. Every block had 2×2 max pooling, ReLU activation function and batch normalization in addition to two 3×3 convolutions. Filter depth increased from 32 to 64, 128, and 256. One global average pooling was followed by a dense layer of 256 units, batch normalization, ReLU, dropout of 0.40 and 1 sigmoid output layer, which is the probability of being malignant. For feature map k , a convolutional activation can be expressed as (3), where W_k is the learned kernel, b_k is its bias, and σ is ReLU.

$$H_k = \sigma(W_k * X + b_k) \quad (3)$$

The output probability was calculated by the sigmoid function in (4). Binary cross-entropy in (5) optimized the model, where y is the true class and p is the malignant probability.

$$p = 1 / (1 + \exp(-z)) \quad (4)$$

$$L_{BCE} = -[y \log(p) + (1-y) \log(1-p)] \quad (5)$$

The same architecture and random seed were used for all branches. For training only augmentation, horizontal flipping, small rotation ($\sim \pm 10^\circ$), $\sim 5\%$ zoom and $\sim 10\%$ translation and contrast adjustment were used. The balanced class weights were 0.9853 for the benign class, and 1.0151 for the malignant class. The parameters of Adam optimization were: initial learning rate of 10^{-3} , L2 regularization of 10^{-4} , batch size of 32 and exactly 15 epochs. ModelCheckpoint kept the model with the lowest validation loss, ReduceLROnPlateau decreased the learning rate by 0.5 after every three epochs without improvement to the smallest value possible of 1×10^{-7} . The early stopping mechanism was intentionally removed to ensure that all three branches have the same amount of training time budget (15 epochs); the same random-seed configuration (42), architecture, augmentation policy, class weighting and optimizer settings were applied to all the branches. The configuration which was implemented is summarized in the Table 3.

Table 3. Implemented CNN and training configuration

Component	Configuration
Input	$224 \times 224 \times 3$ normalized image
Convolutional blocks	Four blocks; two Conv2D-BN-ReLU layers per block
Filters	32, 64, 128, 256
Down-sampling	2×2 max pooling after each block
Classifier	Global average pooling; Dense 256; BN; ReLU; dropout 0.40
Output	One sigmoid neuron: benign (0), malignant (1)
Loss / optimizer	Binary cross-entropy / Adam
Regularization	L2 = 10^{-4} ; dropout = 0.40
Batch / epochs	32 images / exactly 15 epochs
Parameters	1,243,169 total; 1,240,737 trainable

3.5. Independent Evaluation and Statistical Analysis

Validation data was only used for the selection of thresholds for each model. The chosen cut-off value was the one that gave the highest value of Youden's J statistic (sensitivity + specificity - 1). The test set was then evaluated once on the untouched test set. Derivation of accuracy, sensitivity, specificity, precision, negative predictive value, F1 SCORE, Balanced accuracy (BAC), Matthew's correlation coefficient (MCC), Cohen's kappa, ROC-AUC, PR-AUC, Log loss and Brier score were done. Core diagnostic metrics are given in (6)-(10).

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (6)$$

$$\text{Sensitivity} = TP / (TP + FN) \quad (7)$$

$$\text{Specificity} = TN / (TN + FP) \quad (8)$$

$$\text{Precision} = TP / (TP + FP) \quad (9)$$

$$F1 = 2 \times \text{Precision} \times \text{Sensitivity} / (\text{Precision} + \text{Sensitivity}) \quad (10)$$

Uncertainty was calculated by using 2,000 bootstrap re-sampled data and 95 % percentile intervals [19]. Paired errors of the three pairs of models were compared by Exact McNemar test with Holm correction. This design does not assume that correlated predictions are independent and directly verifies the changes on the same test cases due to the preprocessing, not only on the percentages on the whole set of the test's cases [20]. When designing the experiment, no repeated random partitions were performed, nor was the k-fold cross validation used, and all three branches were tested on the same cases not used for training. Repeated-split validation or external validation must be conducted in addition to this independent test cohort. These 2,000-resample bootstrap intervals provide measurements of sampling uncertainty in this test cohort, but are not a substitute for repeated-split or external validation.

3.6. Explainable Artificial Intelligence

Grad-CAM, LIME and SHAP were used to explain the best performing independent-test model. The 12 cases were selected consisting of three correct benign, three correct malignant, three false positive and three false negative. For all 12 cases, Grad-CAM was produced from the last layer of the CNN [16]. LIME employed 500 perturbations, 50 super pixels (SLIC) and up to 10 influential segments [17]. For the latter, SHAP GradientExplainer was trained on 32 cases from the training set, and was used to generate explanations for 8 cases [18]. Explanation entropy, top-20% attribution concentration, central-region concentration and pairwise IoU and Dice agreement were determined. These values are not considered

measures of clinical localization accuracy, but internal consistency measures, as masks were not used by experts.

4. Results and Discussion

4.1. Preprocessing Effectiveness

The representative images illustrate the fact that the intensity variation in the original branch is preserved after applying the raw branch, but after applying the median filter, a smoother background with a preserved general lesion contour is obtained. CLAHE makes the texture separation more obvious in the local area, and enhances the edges of the lesion, while intensifying the remaining structures outside the lesion. This visual trade-off helps to make a multi-metric evaluation instead of an apparent contrast evaluation as in Figure 2, which branches to be selected from.

Standardization and Preprocessing Comparison

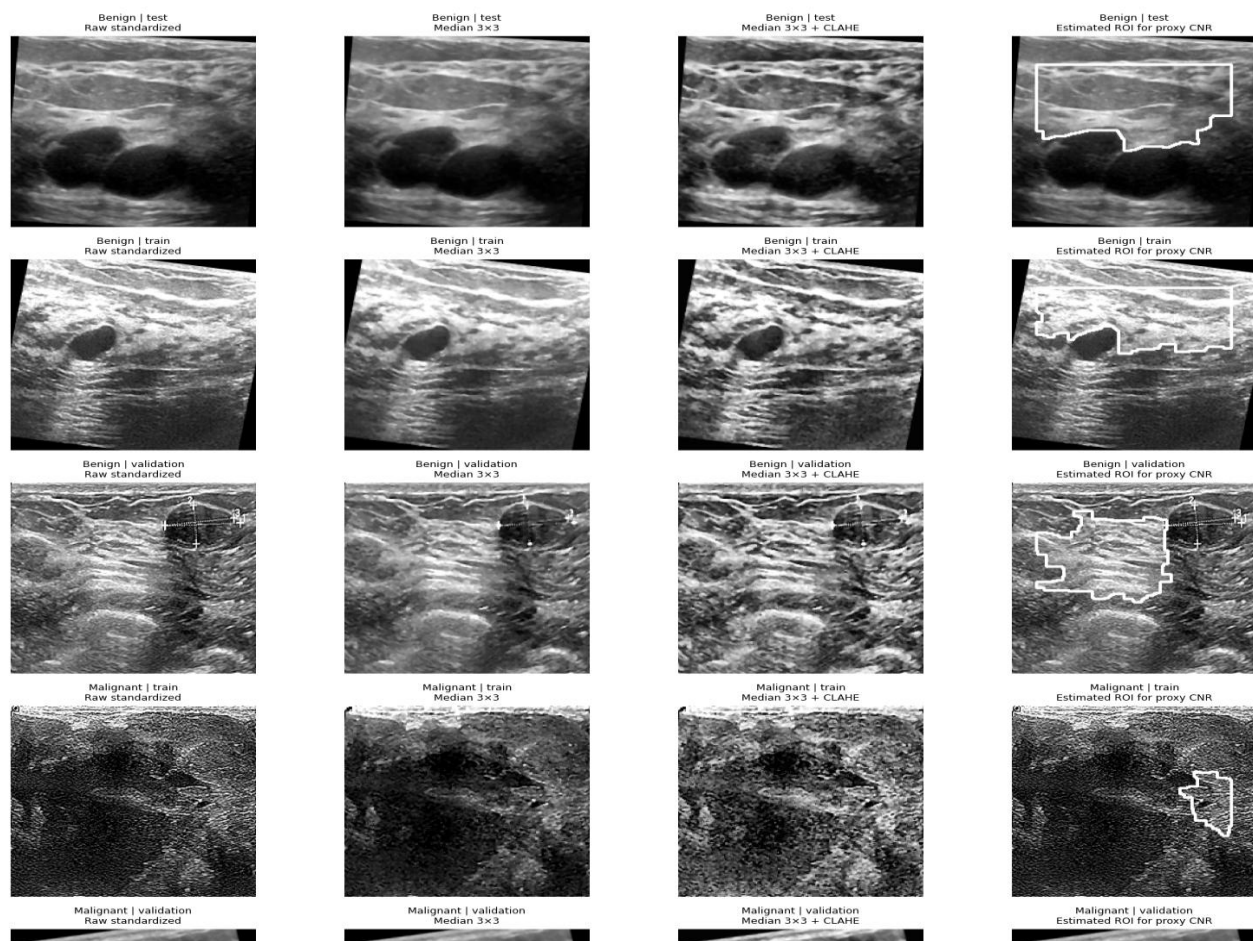


Figure 2. Representative raw standardized, median-filtered, and median-plus-CLAHE images with edge overlays

In all paired images, median filtering led to an increase of mean proxy CNR from 0.6781 to 0.7188 without changing the value of SSIM of 0.8967 with the raw representation. Median-plus-CLAHE yielded the highest entropy and RMS contrast, while having the lowest SSIM, proxy CNR and edge preservation which is an indication of aggressive modification of the image structure. We should not draw the conclusion of superiority of denoising only based on its very large mean ENL, as it was due to the almost homogenous automatically selected regions in some enhanced images. A quality profile has been provided for the entire tree, on branch-by-branch basis, in Table 4.

Table 4. Overall image-quality results for the three preprocessing branches

Branch	SSIM vs raw	Proxy CNR	Entropy	RMS contrast	Edge preservation	ENL
Raw standardized	1.0000	0.6781	7.1955	57.6299	1.0000	0.6936

Median 3×3	0.8967	0.7188	7.1965	55.7346	0.6687	0.7634
Median 3×3 + CLAHE	0.7457	0.5909	7.4416	60.6769	0.6414	2749.5952

The controlled synthetic-speckle experiment provides reference based evidence which is not available in the naturally acquired images. Median restoration improved PSNR from 25.0395 to 27.9313 dB and SSIM from 0.7740 to 0.8214, while increasing proxy CNR from 0.6244 to 0.7134. The addition of CLAHE resulted in a decrease in the fidelity of the reference image to 18.5373 dB PSNR and 0.6522 SSIM. The results of all these distributions are presented in Figure 3 and Table 5.

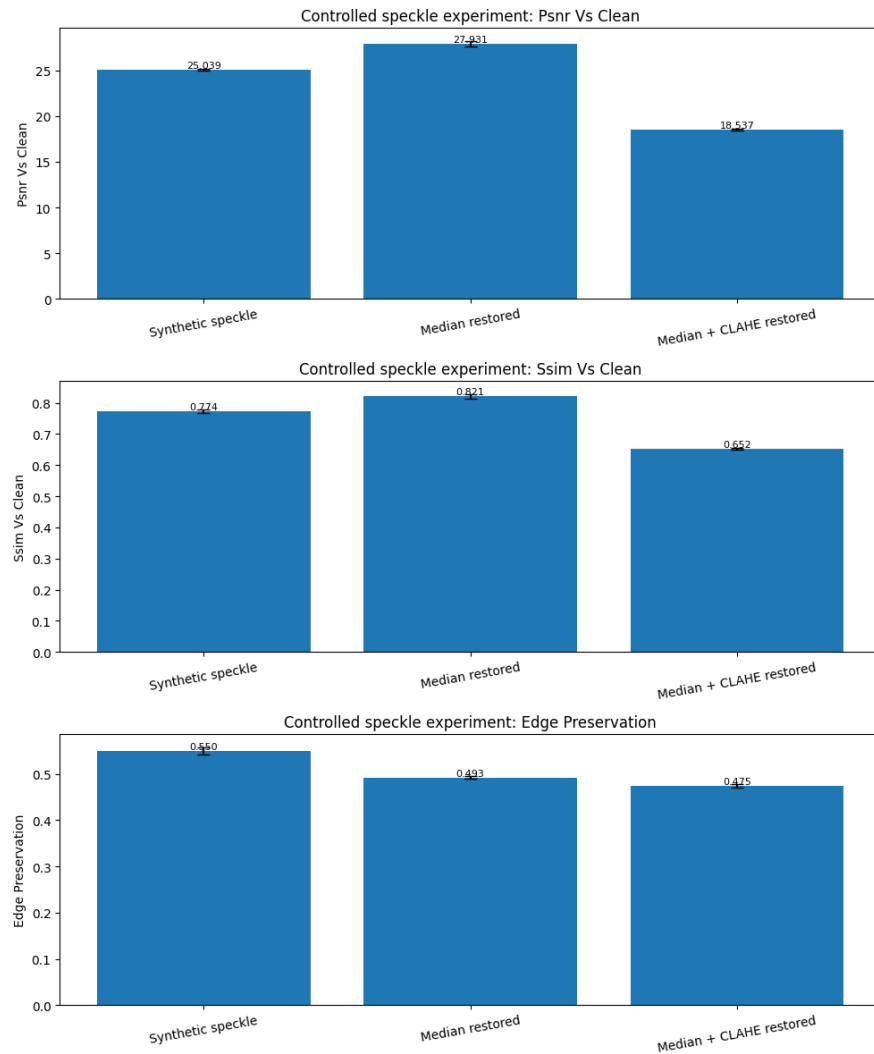


Figure 3. Controlled synthetic-speckle restoration results for PSNR, SSIM, and edge preservation

Table 5. Controlled synthetic-speckle experiment

Condition	PSNR vs clean (dB)	SSIM vs clean	Proxy CNR	Edge preservation
Synthetic speckled	25.0395	0.7740	0.6244	0.5497
Median restored	27.9313	0.8214	0.7134	0.4927
Median + CLAHE restored	18.5373	0.6522	0.5743	0.4746

All of the principal preprocessing contrasts were significant at $p < 0.001$, as determined by paired Wilcoxon tests and Holm correction). The combined evidence suggests that the median filtering branch will reduce the speckle-like variation and improve the contrast related measures while maintaining a considerable amount of structural similarity as compared to the median-plus-CLAHE branch, answering question 1 of the research. The result also shows why the brightness or entropy should not be used to determine the contrast enhancement, but in relation to fidelity and further classification.

4.2. Model Training and Validation

The same budget (15 epochs) was given to all three CNNs. The best value for validation accuracy for the raw branch was 0.93718 while for median filtering was 0.95418 and for median-plus-CLAHE was 0.95565. While CLAHE led to the best validation accuracy, the best validation loss (0.20207) and the highest validation ROC-AUC and PR-AUC (both around 0.99698) were obtained using median filtering. This indicates that even though the CLAHE branch was slightly better at one accuracy threshold the median filtering method was more reliable in the probability separation. Table 6 presents the main results of the validation.

Table 6. Best validation performance during the fixed 15-epoch training schedule

Branch	Best accuracy	Best loss	ROC-AUC	PR-AUC	Best accuracy epoch
Raw standardized	0.93718	0.25034	0.98282	0.98443	15
Median filtered	0.95418	0.20207	0.99698	0.99698	15
Median filtered + CLAHE	0.95565	0.21578	0.99189	0.99185	13

Training curves exhibit a gradual improvement of learning in all branches and more fluctuations in validation loss compared to training loss. This type of variation is in line with a visual related augmentation family's dataset with a high level of overlapping benign-malignant textures. Since early stopping was not employed, the comparison of the branches is not affected by the different training times. As seen in Figure 4, the validation accuracy of the median branch was best and the validation loss was smallest.

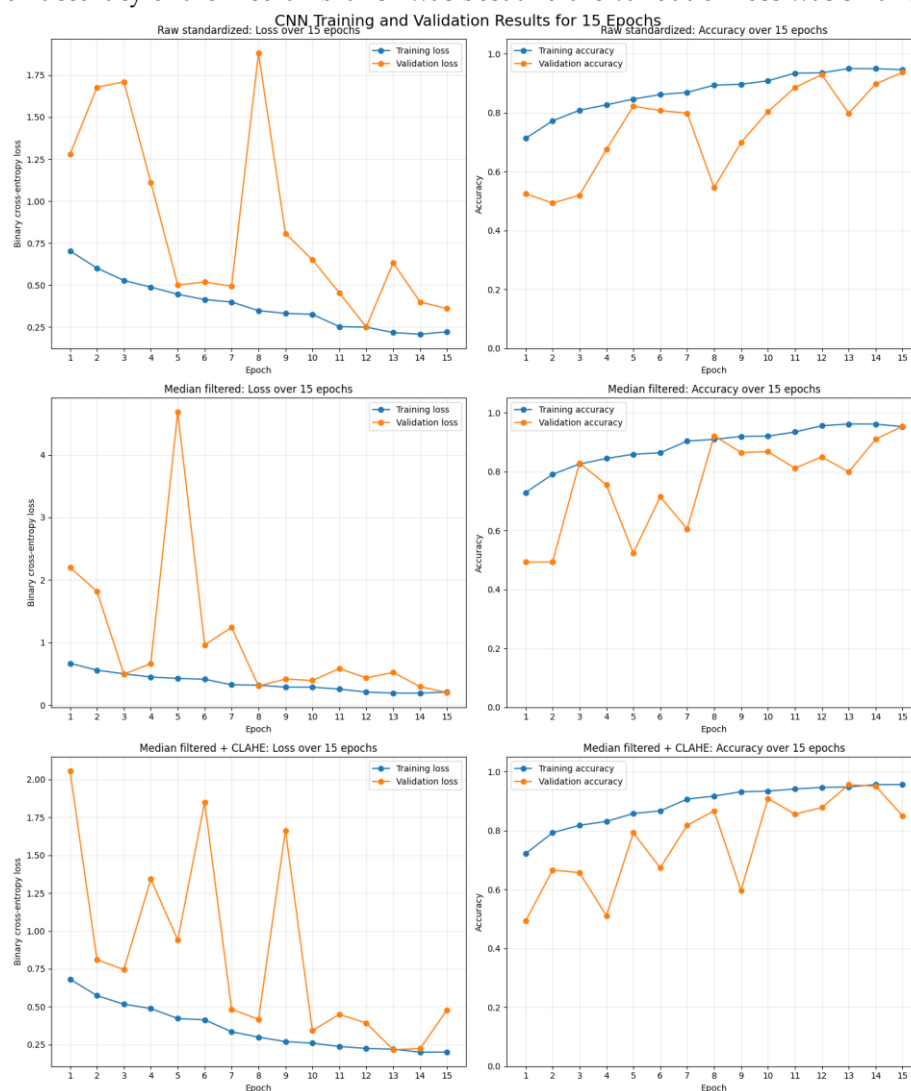


Figure 4. Training and validation loss and accuracy across the three preprocessing branches

4.3. Independent Test Performance

The thresholds obtained from validation were: raw standardized images - 0.53939; median-filtered images - 0.64317 and median-plus-CLAHE images - 0.54205. After correcting these thresholds, they were then used on the 1,353-image independent test set. Median filtering achieved 96.97% accuracy, 98.95% sensitivity, 95.04% specificity, 95.10% precision, and 96.99% F1-score. The gains in accuracy, sensitivity, specificity and F1-score compared to raw inputs were 3.55 percentage points, 6.15, 1.02, and 3.70 respectively. The results of the core diagnostic comparison are presented in Table 7.

Table 7. Core independent test-set classification results

Branch	Accuracy	Sensitivity	Specificity	Precision	F1-score
Raw standardized	93.42%	92.80%	94.02%	93.79%	93.29%
Median filtered	96.97%	98.95%	95.04%	95.10%	96.99%
Median filtered + CLAHE	95.86%	96.10%	95.63%	95.53%	95.82%

These changes can best be clinically interpreted from the confusion matrices. There were 7 false negatives and 34 false positives with the median model correctly classified 660 of 667 malignant cases and 652 of 686 benign cases. The raw model had a 48 false negative and 41 false positive result while the CLAHE model had 26 false negative and 30 false positive result. Therefore, 41 more malignant misses were avoided with the use of median filtering as compared with the raw baseline as shown in Figure 5. Class-wise, the malignant miss rate decreased from 48/667 (7.20%) for raw input to 7/667 (1.05%) after median filtering and 26/667 (3.90%) with median-plus-CLAHE. In the case of benign cases, false positive rates were 41/686 (5.98%), 34/686 (4.96%) and 30/686 (4.37%) for median filter, CLAHE and without any filters respectively, showing that the main benefit of median filtering was the significant reduction in the number of false positives for malignant cases. The sensitivity gain is in agreement with the preprocessing evidence: a 3×3 median filter blocks out isolated speckle-like intensity changes, while preserving the borders of the lesions, which helps reduce nuisance texture that the CNN has to process. CLAHE can also enhance the contrast between local features, while it can also reinforce the image's background features, as seen in the higher number of false negatives compared to the median-only branch. This is an evidence-based mechanistic explanation, not an evidence-based causal explanation.

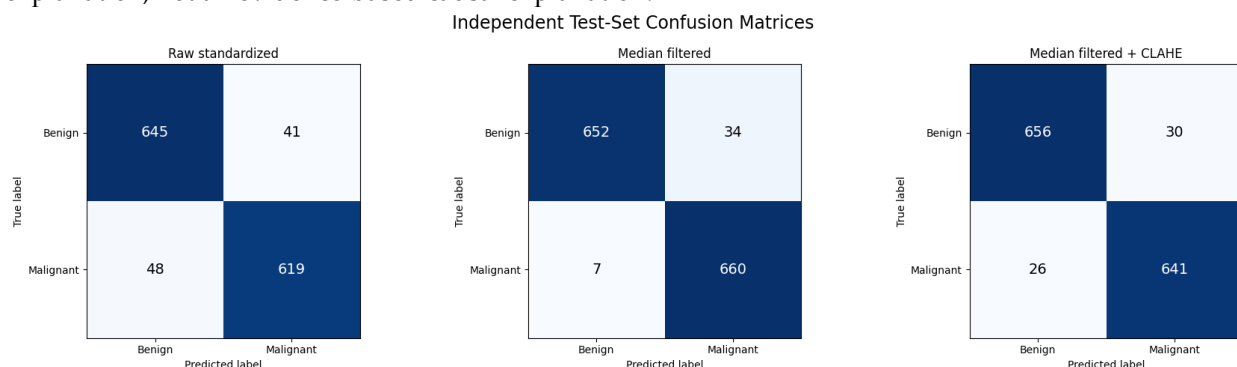


Figure 5. Independent test-set confusion matrices for raw, median, and median-plus-CLAHE models

This is confirmed by threshold independent measures. Median filtering achieved ROC-AUC 0.99623, PR-AUC 0.99608, balanced accuracy 0.96997, and MCC 0.94016. Its log loss of 0.13027 was lower than the raw model's 0.16618. Median-plus-CLAHE had the lowest Brier score (0.03593) which showed competitive calibration of the method, while its discrimination and error profile were lower as compared to median filtering. The ROC curves show that the operating range of the median branch dominates much of the operating range as shown in Figure 6.

Resampling of the final median model in the bootstrap analysis indicated that the model was stable. Its 95% intervals were 0.96083–0.97857 for accuracy, 0.98085–0.99693 for sensitivity, 0.93363–0.96547 for specificity, 0.96081–0.97859 for F1-score, and 0.99446–0.99782 for ROC-AUC. Specificity was the comparatively less component, suggesting that there is room for further improvement of the test in terms of false-positive benign cases. Summary of the confidence intervals in figure 7. These were calculated from 2,000 resampling of the single held-out test split that explicitly measure uncertainty with respect to the accuracy, sensitivity, specificity, F1-score, and ROC-AUC on that test split.

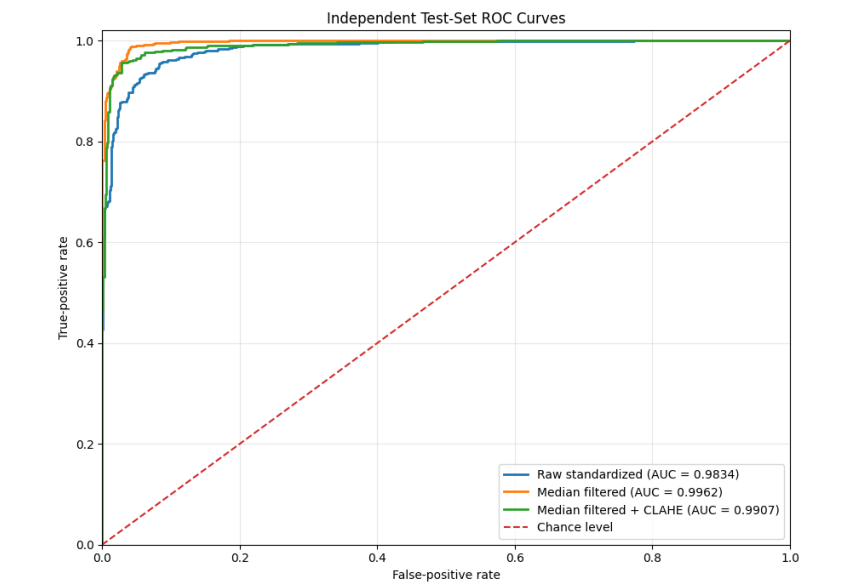


Figure 6. Independent test-set receiver operating characteristic curves

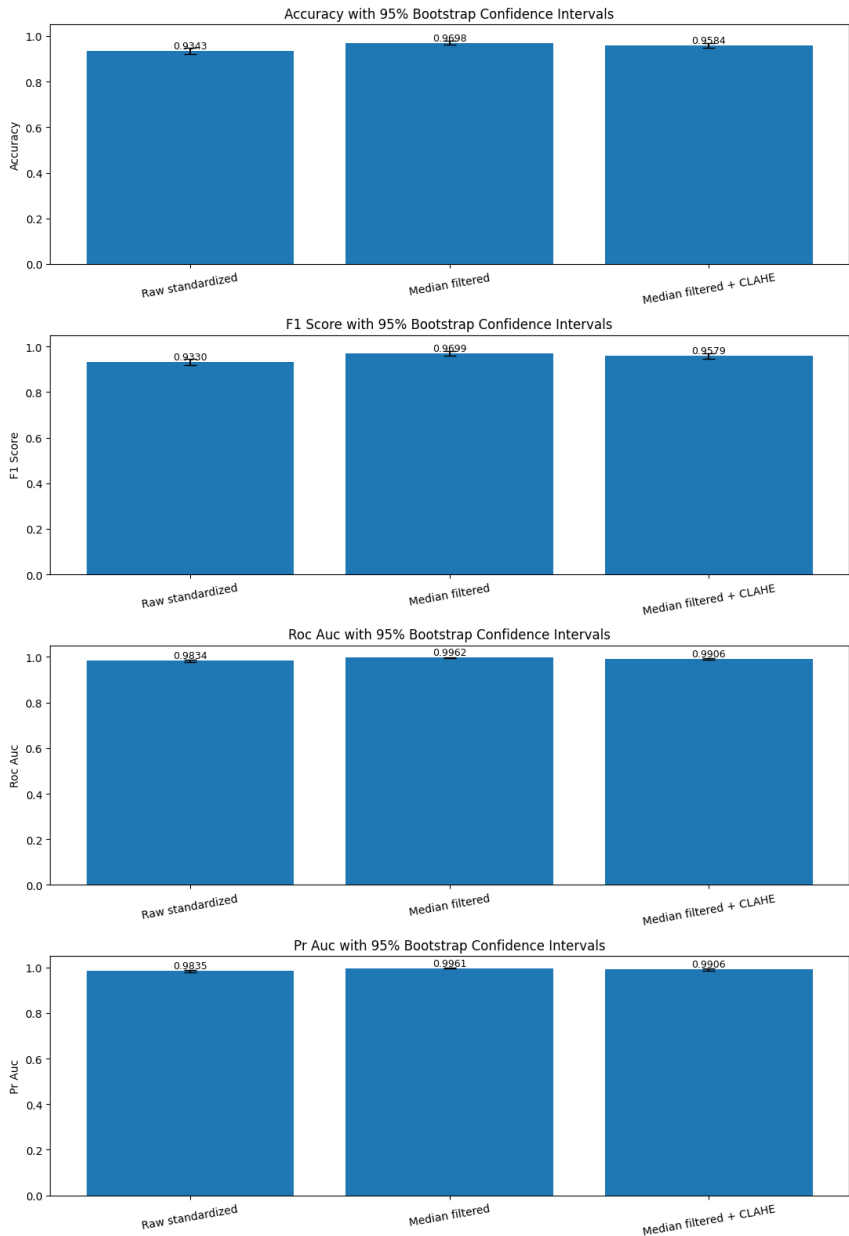


Figure 7. Bootstrap confidence intervals for principal independent test metrics

Enhanced branches were found to commit fewer errors with both tests being exact paired McNemar tests compared to the raw model. However, raw versus median and raw versus median-plus-CLAHE were still significant throughout the Holm correction procedure (adjusted $p = 0.000008$, and adjusted $p = 0.004066$, respectively). Even though the point estimates were more favorable for the median filtering, there was no significant difference between median and median-plus-CLAHE (adjusted $p = 0.137368$). The number of paired errors is reported in Table 8.

Table 8. Exact pairwise McNemar tests with Holm correction

Model comparison	A correct / B wrong	A wrong / B correct	Holm-adjusted p	Significant
Raw vs median	28	76	0.000008	Yes
Raw vs median + CLAHE	38	71	0.004066	Yes
Median vs median + CLAHE	52	37	0.137368	No

The following results provide answer to the second research question. A statistically significant improvement was obtained with the preprocessed standardized inputs with respect to the raw standardized inputs under a matched architecture and test set. The median-only branch had the best clinical error balance as it almost had a 99% sensitivity and a false negative count, which is 1/7 of the count of the raw model. The negative result for the median vs. the CLAHE approach needs to be stated with caution: although the median filter gave the best performance in this experiment, it cannot be concluded that it is better than all the CLAHE versions based on this experiment; this test should be conducted on other datasets.

4.4. Explainability Results

The final model (after median filtering) was analyzed by Grad-CAM, LIME and SHAP on correctly classified benign, correctly classified malignant, false-positive and false-negative predictions. The resulting maps were broad class-specific activations (Grad-CAM), localized influential super pixels (LIME) and localized positive and negative attribution with respect to the background distribution (SHAP). The methods often focused on overlapping areas of the lesion and boundary in correct cases. In this false positive case (illustrated), attention was paid to the adjacent texture, implying that this irregular border contrast was a factor in the malignant diagnosis. Some representative explanations are provided in Figure 8.

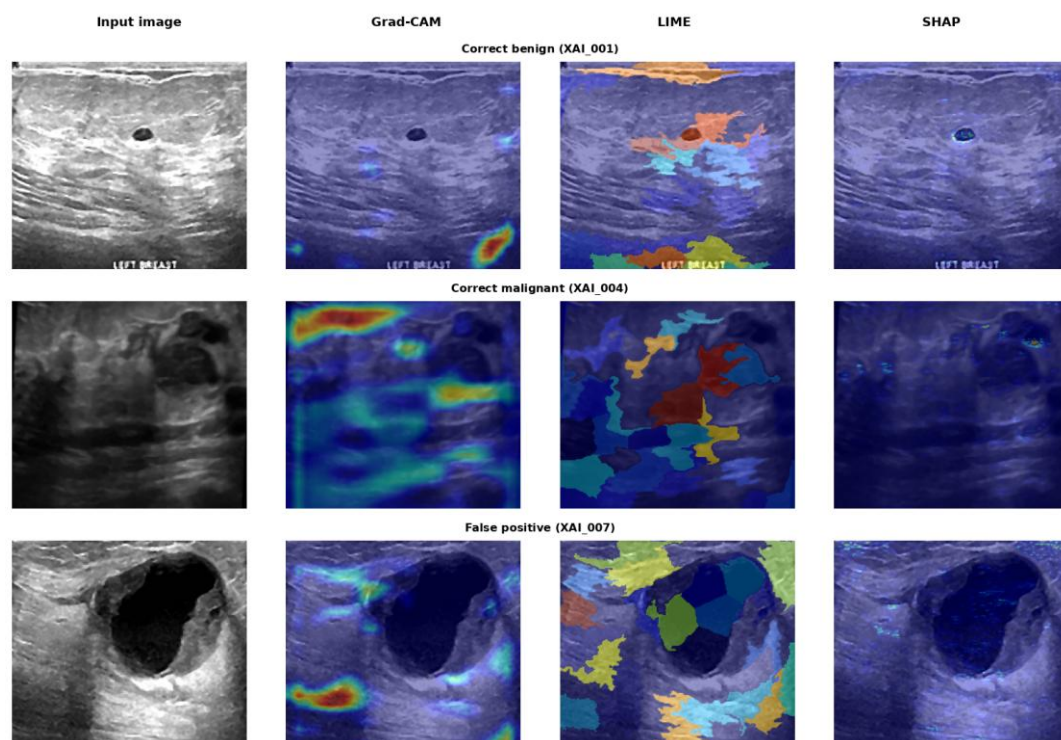


Figure 8. Representative Grad-CAM, LIME, and SHAP explanations for the median-filtered model

In terms of absolute numbers, the mean concentration for top 20% for Grad-CAM, LIME and SHAP were 0.7156, 0.6770 and 0.5682, respectively. This ranking doesn't mean that one technique is more accurate, it is simply reflecting of the differences in construction. The values of Mean IoU for Grad-CAM and LIME, LIME and SHAP and SHAP and Grad-CAM were 0.1666, 0.1100 and 0.1373 respectively. The agreement was generally higher in correctly classified cases than in the explained cases of false positives, for instance, the average of the IoU of Grad-CAM was ~ 0.1307 for correct cases and ~ 0.0479 for the false-positive explained cases. Figure 9 shows the pattern of agreement between pairs of people.

The small agreement values provide a response to the third research question indicating complementary evidence from the methods. One heatmap might not include features that are shown by another explanation system. Clinical localizations, on the other hand, cannot be made due to the lack of expert lesion masks. It is logically reasonable to conclude that XAI made models more transparent, unveiled patterns of failure, and helped to audit models but not that all the highlighted parts were pathologically correct. The overlap is also intuitively low as the following explains: Grad-CAM is a gradient-based explanation on a spatially coarse level; LIME relies on superpixel segmentation and perturbation sampling; SHAP relies on the background distribution chosen and the signed attribution values. Therefore, low IoU is not to be interpreted as explanation failure, since it is a consequence of different spatial resolutions and attribution semantics. Since there are no lesion masks available for the experts, the study cannot determine which of the explanations is the most anatomically faithful, and the maps can only be used as models for auditing.

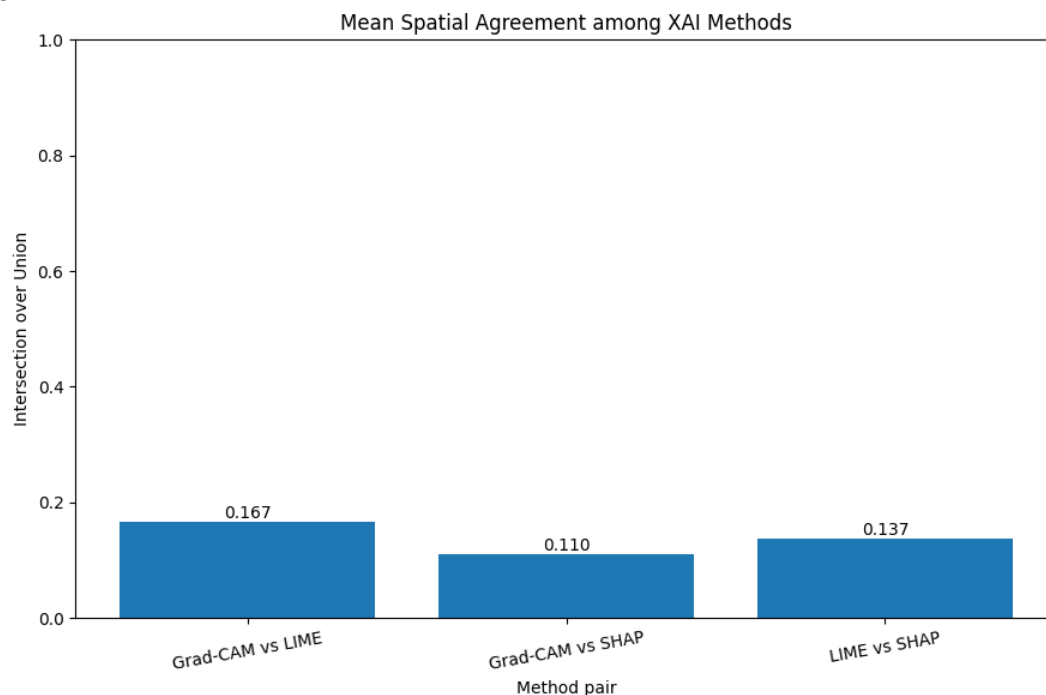


Figure 9. Mean spatial agreement among Grad-CAM, LIME, and SHAP explanation maps

4.5. Comparison with State-of-the-Art Methods

The direct comparison among publications is possibly misleading due to the different versions of the datasets, number of classes, patient independence, augmentation, external validation and cross validation design. The present model, however, performs well compared to the last few public-dataset results: it is better than the meta-learning BUSI ensemble which reported approximately 90% accuracy [3], the BUSI accuracy reported in work [12] with DNBCD (89.87%) and the DRM-Net (96.71%) which approaches or exceeds the discrimination reported in transformer ensembles [5]. The more robust point estimate needs to be read in conjunction with the binary point estimate from the pre-augmented archive for the present study, and the family-aware split. A comparison of protocol aware is summarized in Table 9.

Table 9. Protocol-aware comparison with representative recent breast ultrasound studies

Study	Dataset / Task	Method	Reported result	Comparison limitation
Ali et al. [3]	BUSI; binary	Meta-learning CNN ensemble	≈90% accuracy	Ensemble; split comparability unclear
Alotaibi et al. [4]	BUSI + 2 other datasets	Filtering, ROI highlighting, RGB fusion, VGG19	Preprocessing improved predictions	Different cross-validation and augmentation
Zhou and Mosadegh [5]	Curated BUSI; binary	ViT ensemble / distilled ViT	AUC 0.977 / 0.972	Five-fold CV; architecture-focused
Islam et al. [8]	Two BUS datasets	EDCNN + U-Net + Grad-CAM	87.82% / 85.69% accuracy	Segmentation-dependent
Yan et al. [9]	Private multimodal cohort	Interpretable prototype network	AUC 0.902	Clinically strong, but private data and multimodality
Saini et al. [11]	Two public + in-house	2D-VMD guided attention CNN	98%, 93%, 89% accuracy	Different datasets and validation schemes
Alom et al. [12]	BUSI; three classes	DenseNet121 + Grad-CAM	89.87% accuracy	Three-class task
Ganie et al. [13]	BUS ultrasound	Hybrid DRM-Net + CAM	96.71% accuracy; 97% recall; AUC 0.99	Stacked transfer-learning model
Present study	9,016-image Kaggle archive; binary	Family-aware split; median filtering; fixed CNN; 3 XAI methods	96.97% accuracy; 98.95% sensitivity; AUC 0.99623	Single archive; no masks or external clinical cohort

The difference is thus not so much that these structures are the ultimate in architectural achievement, but methodological. The study makes the connection between family-aware data handling, paired preprocessing quality, a fixed CNN experiment, independent thresholding, bootstrap uncertainty, paired error testing, and three-method XAI. This is a controlled design which is the reason for the improvement in the performance of the median branch and identifies further weaknesses. The main drawback is that the data were derived from a single public database, and did not include patient ID, lesion masks, images from other hospitals or a prospective reader study. The present design process is simpler than multi-stage pre-processing and transfer-learning approach of Alotaibi et al. [4] and trainable speckle-suppression approach of Wang et al. [7] since the former approach involves more numerous additional components that obfuscate the role of the speckle-suppression method. Transformer, decomposition-guided, and stacked transfer-learning approaches [5, 11, 13] have been shown to perform well, but they are based on different datasets, validation schemes, external/in-house cohorts, or much more complex models. Yan et al. [9] on the other hand has more clinical evidence in biopsies based multimodal data and radiologist assessment. Therefore, the results of the present study must be regarded as competitive to the mentioned studies and more favorable in the controlled preprocessing isolation. Thus, the present study should be interpreted as competitive in predictive performance and more favorable in controlled pre-processing isolation, but not as proof of clinical superiority to these heterogeneous studies. The extent of clinical generalizability is restricted by the sole Kaggle archive, the lack of scanner/site and demographic metadata, inferred (not patient-verified) familial relationships, and the absence of an external hospital cohort and of a prospective radiologist reader study. The estimated counts reported are therefore a representation of the curated collection, and it is not necessarily representative of counts at other institutions, on other devices, in other populations, or under different acquisition systems.

5. Conclusion

This paper proposed a leakage aware classification framework for breast ultrasound images which separated out the effect of image preprocessing with the same CNN and evaluation system. Median filtering of the raw data was the most effective of three different inputs (raw data, raw data standardized with a 3×3 median filter, and median + CLAHE) in terms of denoising, structural fidelity, and diagnostic performance. It provided the highest proxy CNR, the best controlled restoration values (27.9313 dB PSNR and 0.8214 SSIM), and the high independent test accuracy (96.97%), sensitivity (98.95%), specificity (95.04%), F1-score (96.99%) and ROC-AUC (0.99623). There was a clinically and statistically significant reduction in the number of false negatives from 48 to 7 and a significant raw-versus-median McNemar result. Additionally, Grad-CAM, LIME, and SHAP were able to explain their predictions correctly and incorrectly, and they achieved fairly low spatial overlap, which showed that considering more than one explanation technique was useful.

The proposed framework is more powerful than accuracy-only studies as it merges family aware splitting, matched branch comparison, broad diagnostic metrics, confidence intervals, paired significance testing and multi-method XAI. It still has limitations: it utilizes a single publicly available archive, it infers source families from the images, it classifies only the binary classes, it does not have lesion masks, it only uses a CNN model with 15 epochs, and there are no external clinical validations. Future work will include the extension of the pipeline to 3D ultrasound, prospective acquisition of multi-institutional cohorts of subjects, radiologist-drawn lesion contours, real-time uncertainty-aware deployment, semi-supervised and federated learning, and cross-modality evaluation with mammography and MRI. Clinical claims are further restricted by the possibility of a public pre-augmented archive, unverified patient-level independence and the lack of external or prospective evaluation of the archive as well as the lack of scanner and demographic metadata. Patient cohorts identified by patients, device/site diversity, radiologist review, and prospective testing of workflow should therefore be considered first in future validation studies before clinical deployment. Clinical claims are further restricted by the possibility of a public pre-augmented archive, unverified patient-level independence and the lack of external or prospective evaluation of the archive as well as the lack of scanner and demographic metadata. Patient cohorts identified by patients, device/site diversity, radiologist review, and prospective testing of workflow should therefore be considered first in future validation studies before clinical deployment.

References

1. W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, p. 104863, 2020, doi: 10.1016/j.dib.2019.104863.
2. H. Kim, J. Park, H. Lee, G. Im, J. Lee, K.-B. Lee, and H. J. Lee, "Classification for breast ultrasound using convolutional neural network with multiple time-domain feature maps," *Applied Sciences*, vol. 11, no. 21, p. 10216, 2021, doi: 10.3390/app112110216.
3. M. D. Ali et al., "Breast cancer classification through meta-learning ensemble technique using convolution neural networks," *Diagnostics*, vol. 13, no. 13, p. 2242, 2023, doi: 10.3390/diagnostics13132242.
4. M. Alotaibi, A. Aljouie, N. Alluhaidan, W. Qureshi, H. Almatar, R. Alduhayan, B. Alsomaie, and A. Almazroa, "Breast cancer classification based on convolutional neural network and image fusion approaches using ultrasound images," *Heliyon*, vol. 9, no. 11, e22406, 2023, doi: 10.1016/j.heliyon.2023.e22406.
5. G. Zhou and B. Mosadegh, "Distilling knowledge from an ensemble of vision transformers for improved classification of breast ultrasound," *Academic Radiology*, vol. 31, no. 1, pp. 104–120, 2024, doi: 10.1016/j.acra.2023.08.006.
6. A. A. Hayum, J. Jaya, R. Sivakumar, and B. Paulchamy, "An efficient breast cancer classification model using bilateral filtering and fuzzy convolutional neural network," *Scientific Reports*, vol. 14, p. 6290, 2024, doi: 10.1038/s41598-024-56698-8.
7. R. Wang, X. Liu, and G. Tan, "Coupling speckle noise suppression with image classification for deep-learning-aided ultrasound diagnosis," *Physics in Medicine & Biology*, vol. 69, no. 6, 2024, doi: 10.1088/1361-6560/ad29bb.
8. M. R. Islam et al., "Enhancing breast cancer segmentation and classification: An ensemble deep convolutional neural network and U-Net approach on ultrasound images," *Machine Learning with Applications*, vol. 16, p. 100555, 2024, doi: 10.1016/j.mlwa.2024.100555.
9. L. Yan et al., "A domain knowledge-based interpretable deep learning system for improving clinical breast ultrasound diagnosis," *Communications Medicine*, vol. 4, p. 90, 2024, doi: 10.1038/s43856-024-00518-7.
10. C. Aumente-Maestro, J. Díez, and B. Remeseiro, "A multi-task framework for breast cancer segmentation and classification in ultrasound imaging," *Computer Methods and Programs in Biomedicine*, vol. 260, p. 108540, 2025, doi: 10.1016/j.cmpb.2024.108540.
11. M. Saini, S. Hassanzadeh, B. Musa, M. Fatemi, and A. Alizad, "Variational mode directed deep learning framework for breast lesion classification using ultrasound imaging," *Scientific Reports*, vol. 15, p. 14300, 2025, doi: 10.1038/s41598-025-99009-5.
12. M. R. Alom et al., "An explainable AI-driven deep neural network for accurate breast cancer detection from histopathological and ultrasound images," *Scientific Reports*, vol. 15, p. 17531, 2025, doi: 10.1038/s41598-025-97718-5.
13. S. M. Ganie et al., "Explainable AI based hybrid DRM-Net transfer learning model for breast cancer detection and classification using ultrasound images," *Scientific Reports*, vol. 15, p. 44170, 2025, doi: 10.1038/s41598-025-27934-6.
14. M. Fatima et al., "An adaptive deep learning approach based on InBNFus and CNNDen-GRU networks for breast cancer and maternal fetal classification using ultrasound images," *Scientific Reports*, vol. 15, p. 22027, 2025, doi: 10.1038/s41598-025-03402-z.
15. S. Saharan et al., "A deep learning and explainable artificial intelligence based scheme for breast cancer detection," *Scientific Reports*, vol. 15, p. 32125, 2025, doi: 10.1038/s41598-024-80535-7.
16. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision*, 2017, pp. 618–626.
17. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
18. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
19. B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall/CRC, 1994.
20. Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, pp. 153–157, 1947.
21. S. M. Pizer et al., "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355–368, 1987.
22. A. Heikal, A. El-Ghamry, S. Elmougy, and M. Z. Rashad, "Fine tuning deep learning models for breast tumor classification," *Scientific Reports*, vol. 14, p. 10753, 2024, doi: 10.1038/s41598-024-60245-w.