XAI for a breast lesion classifier

Sofía Pedrós Tobaruela

December 26, 2025

1 Problem

The problem presented in this work is the classification of breast lesions to distinguish between benign and malignant cases. This type of classifiers is used to assist radiologists and oncologists in early breast cancer detection and diagnosis.

The primary stakeholders in this scenario are the radiologists and oncologists, who rely on the model's output to support diagnostic decisions, and patients, whose treatment outcomes may be influenced by these decisions.

The model's explanations provide transparency that can support doctors in confirming whether the automated decisions are reasonable or not. Explainability in this context is crucial because incorrect predictions can lead to delayed treatment of unnecessary invasive procedures. Additionally, if the model's predictions are usable and accurate, the medical institution could delay the immediate intervention of a radiologist, helping reduce their workload and optimize medical resources.

2 Dataset

The dataset used in this work is the *Breast-Lesions-USG* dataset [2], which is publicly available at `https://www.cancerimagingarchive.net/collection/breast-lesions-usg/`.

The dataset has 256 breast ultrasound scans collected from 256 patients, including benign, normal and malignant cases. All cases were confirmed by follow-up care or biopsy result. Each scan was manually annotated and labeled by a radiologist experienced in breast ultrasound examination.

The dataset includes: image, segmentation (pixel-wise mask with the tumor region or abnormal areas) and tabular clinical data. The mask follows the BI-RADS standard, including category and descriptors. The BI-RADS standard is a system doctors use to classify breast images in seven categories: insufficient data, normal, benign, probably benign, suspicious, trustworthy indicator of malignant and malignant confirmed by biopsy [1].

This rich annotation structure enables the dataset to be used for classification, segmentation, and explainable AI tasks.

Several preprocessing steps were applied prior to model training. Tabular data preprocessing included the removal of non-informative identifiers, conversion of categorical variables using either one-hot encoding or label encoding, and mapping of binary clinical attributes (e.g., yes/no) to numerical values. Missing or unavailable values were converted to zero. Continuous numerical features were standardized to zero mean and unit variance, while binary features were left unchanged. The final dataset was split into training, validation, and test subsets (80%, 10% and 10%) using a fixed random seed to ensure reproducibility.

3 Tabular model

I have trained a multilayer perceptron (MLP) model to classify the lesions using tabular data obtained from the ultrasound images.

In an initial approach, all available tabular features were used (except for the ones with the BI-RADS results, that could give away the result immediately). Two encoding strategies for categorical variables were evaluated: one-hot encoding and label encoding. One-hot encoding resulted in artificial and less interpretable feature representations (e.g., *Signs_yes*, *Signs_no*). Therefore, label encoding was selected, assigning integer values to categorical variables (e.g., shape: oval=0, round=1, irregular=2). No significant difference in performance was observed between the two approaches.

The MLP architecture consisted of two hidden layers with 128 and 256 neurons, ReLU activations, and dropout regularization. The model was trained for 40 epochs using a learning rate of 10^{-4} and a batch size of 8.

The initial model achieved an accuracy of 89.71% on the training set, 84.62% on the validation set, and 88.46% on the test set.

4 Global explanations

To understand the model's behavior and identify irrelevant or redundant features I generated global explanations using SHAP and permutation importance. Both methods are model-agnostic and provide complementary perspectives on feature importance.

The global explanations revealed that several features showed low importance across both methods, as it can be seen in Figure 1. The features *posterior features*, *tissue composition*, *skin thickening*, and *symptoms* are not relevant to the model's prediction, so they were removed in a second training iteration.

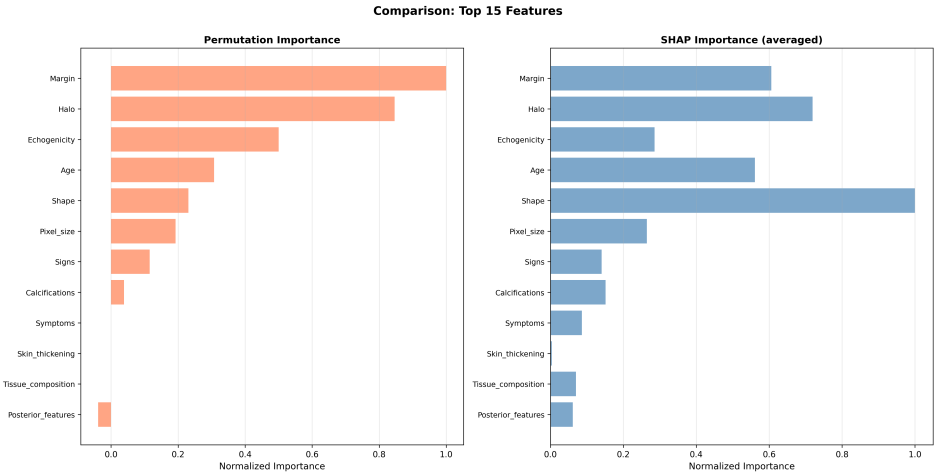


Figure 1: Shap and permutation importance global explanation for the first trained model

Although the *Age* feature was identified as important by both explainability methods, it was also removed based on expert reasoning. In realistic image only scenarios, patient age may not be available to the model at inference time. To avoid reliance on unavailable information, this feature was excluded.

This first model achieved an accuracy of 89.71% on the training set, 84.62% on the validation set, and 88.46% on the test set.

With these features removed, I trained a second model. Despite the reduction in input features, the model maintained the same performance, achieving 88.46% accuracy on the test set, which demonstrates that the features that were removed were not very relevant to the final prediction, so by removing them we can train a model that is equally useful but simpler to interpret. Even though age was more relevant than the other ones, surprisingly removing it did not have a negative impact on the performance. The results for the global explanations of this approach can be found on Figure 2. This Figure shows how the Halo, Echogenicity, Shape and Margin are the most important features in general. It appears that there is some discrepancy between the two explanations, but it looks like all the features that are left are important to some extent.

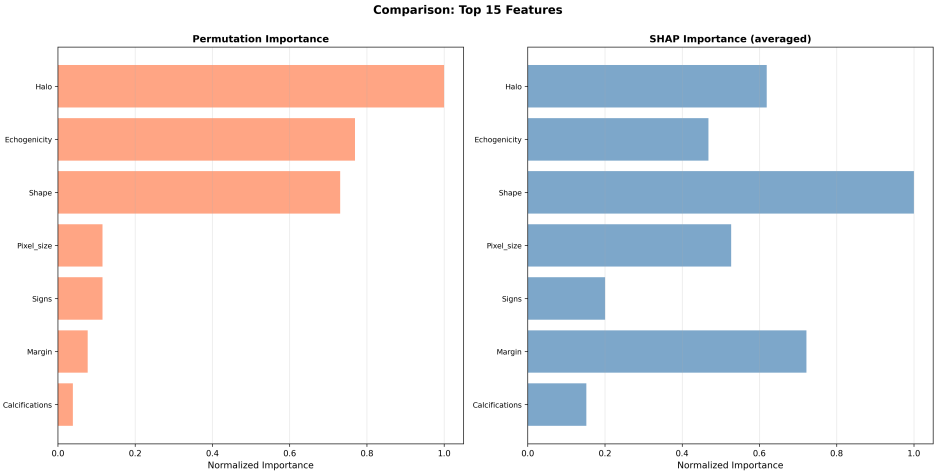


Figure 2: Shap and permutation importance global explanation for the first trained model

In this second model, I also looked at how the predictions were made. In medical diagnosis, false negatives (classifying a malignant lesion as benign) are significantly more harmful than false positives. While false positives can be corrected through follow-up examinations, false negatives may delay treatment and negatively affect patient outcomes.

To address this, I introduced a third training approach by adding weights to the loss to penalize malignant tumors being classified as bening. Several penalty coefficients were evaluated, and a value of 1.5 provided the best trade-off between overall accuracy and false negative reduction. The classes are unbalanced, as it can be seen on Figure 3, so the loss weight also takes that into account.

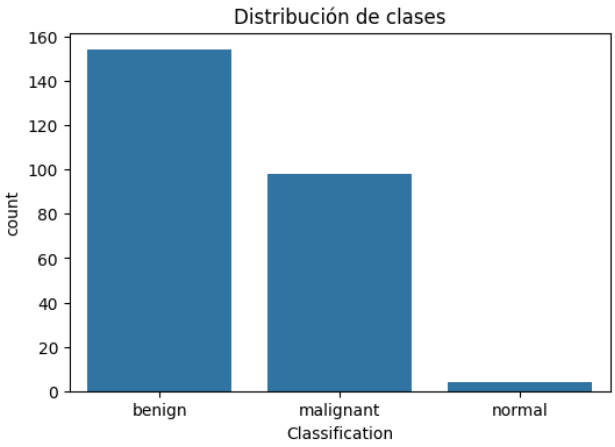


Figure 3: Class distribution

The confusion model for the second model can be found on Figure 4 and the one for the third model on

Figure 5. From these figures we can see that the third approach does not produce as much false negatives, so it is reducing the risk classifying malignant tumors as benign, while maintaining accuracy.

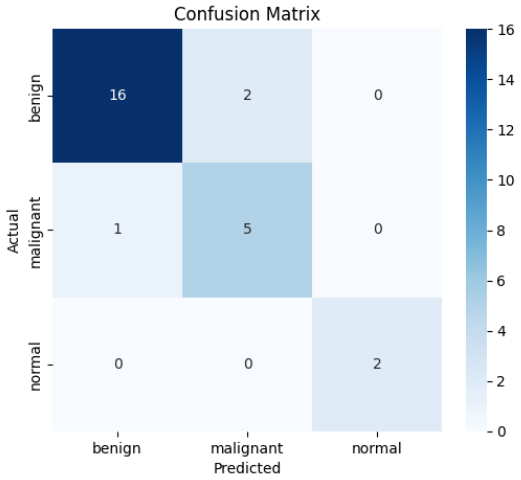


Figure 4: Confusion matrix for the second approach

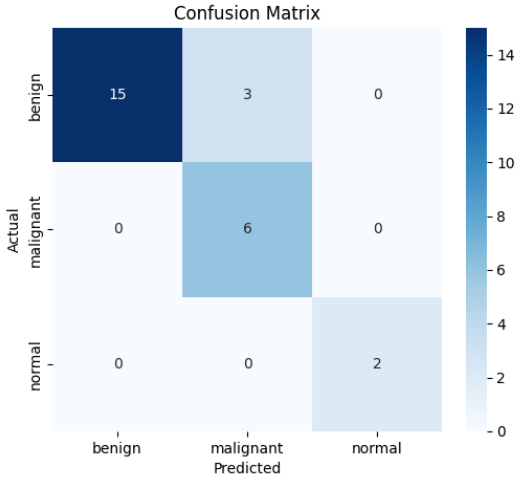
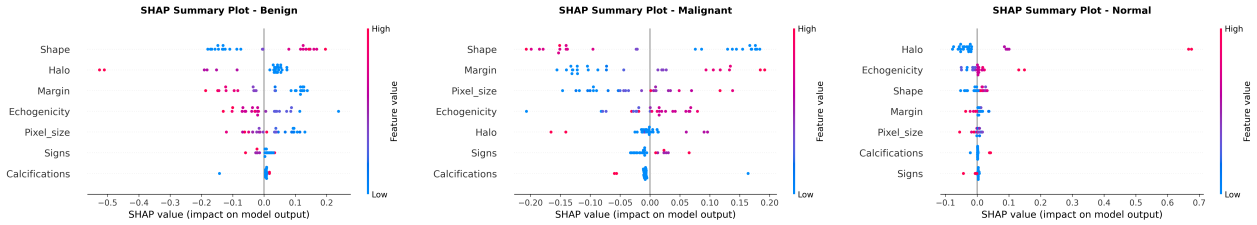


Figure 5: Confusion matrix for the third approach

The third (and final) approach was trained for 40 epochs with a $1e-4$ learning rate, batch size 8, 0.2 dropout, and two hidden layers of shapes 128 and 256 respectively. It achieved an accuracy of 89.71% on the training set, 84.62% on the validation set, and 88.46% on the test set. This indicates that cost-sensitive learning improved clinical safety without sacrificing overall performance.

SHAP summary plots were also generated for each class to analyze class-dependent feature contributions, and the plots can be found on Figure 6. Features such as lesion shape and margin exhibited opposite effects for benign and malignant predictions. For example, irregular shapes tended to push predictions towards malignancy, while regular shapes favored benign classifications.

These class-specific explanations were consistent with medical knowledge and aligned with the global feature importance results, reinforcing confidence in the learned decision patterns.



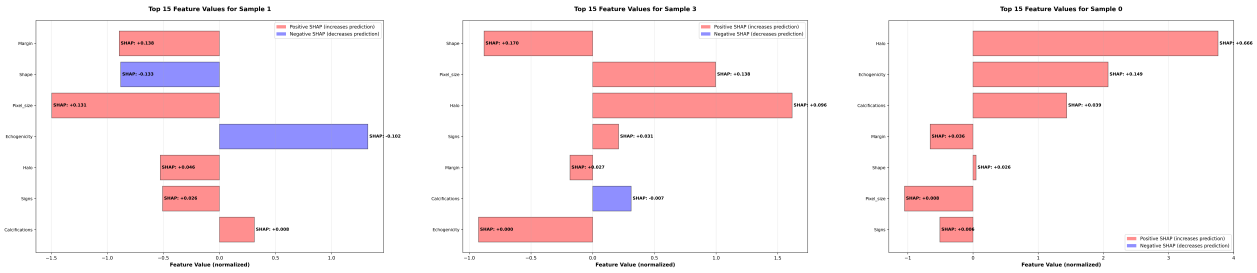
(a) Benign class (b) Malignant class (c) Normal class

Figure 6: SHAP summary plots for the benign, malignant, and normal classes. The figures illustrate the class-specific contribution of features to the predictions.

5 Local explanations

To further assess interpretability, individual predictions were explained using SHAP with a baseline of 100 randomly selected samples. For each class, representative samples were analyzed.

Local explanations largely aligned with global trends; however, some features exhibited strong influence in individual cases despite lower global importance. This behavior is expected, as SHAP provides local explanations that capture instance-specific decision patterns.



(a) Benign class (b) Malignant class (c) Normal class

Figure 7: SHAP summary plots for a benign, malignant, and normal sample.

6 Sanity Checks

To evaluate the robustness of the explainability methods, several sanity checks were conducted. The goal was to assess whether the feature importance rankings provided by permutation importance and SHAP were meaningful. In particular, I removed and added noise (in separate experiments) to the most important features according to SHAP and analyzed the resulting changes in model performance. In addition, random features were removed to serve as a baseline comparison.

First, progressively removing the most important features in descending order of SHAP importance produced the results shown in Figure 8. As expected, the model accuracy generally decreases as the most relevant features are removed, indicating that the SHAP ranking captures meaningful information. A slight increase in accuracy is observed when removing features ranked 6 and 7. This behavior is likely due to the implementation choice of replacing removed features with their mean values, which may still carry informative signal. Nevertheless, the overall monotonic degradation trend supports the robustness of SHAP feature importance.

Next, random features were removed (also replaced by their mean values). The results, shown in Figure 9, demonstrate that removing the top five most important features leads to a substantial drop in accuracy, whereas removing five random features has a significantly smaller impact. This contrast further validates the relevance of the features identified by SHAP.

Finally, noise was added to selected features to further assess explanation robustness. The results are presented in Figure 10, which illustrates how accuracy, prediction changes, and class probability variations evolve as noise increases. An unexpected increase in accuracy is observed at moderate noise levels (noise magnitude of 0.5), likely due to stochastic effects of noise pushing feature values toward more favorable regions of the decision space. However, when substantial noise is added to important features, model accuracy drops sharply. Changes in predictions and probability distributions show more consistent behavior because perturbing important features leads to larger prediction shifts and probability variations than perturbing randomly selected features.

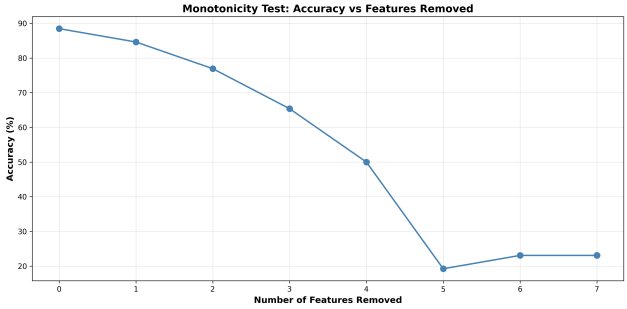


Figure 8: Monotonicity test showing the effect of progressively removing the most important features according to SHAP on model accuracy.

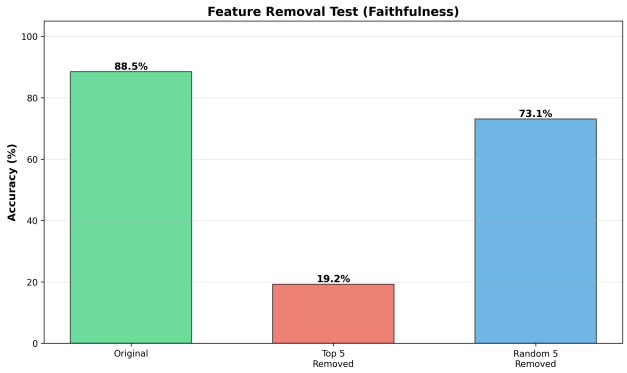


Figure 9: Comparison between removing the most important features and removing random features.

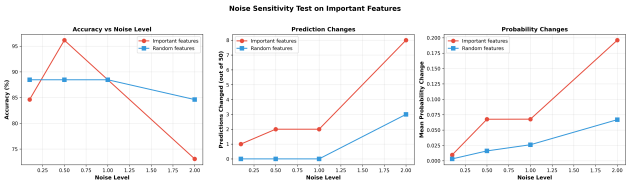


Figure 10: Effect of adding noise to important and random features on model accuracy, prediction stability, and class probability distributions.

References

[1] Breastcancer.org. Sistema de datos e informes de imágenes mamarias (bi-rads), 2025. Revisado por Kevin Fox, MD. Actualizado el 27 de julio de 2025.

[2] Agnieszka Pawłowska, Aleksandra Ćwierz-Pieńkowska, Anna Domalik, et al. Curated benchmark dataset for ultrasound based breast lesion analysis. *Scientific Data*, 11:148, 2024.