Interpretable breast cancer classification using CNNs on mammographic images

Ann-Kristin Balve AKDB3@CAM.AC.UK
University of Cambridge, United Kingdom

Peter Hendrix P.H.G.HENDRIX@TILBURGUNIVERSITY.EDU
Tilburg University, The Netherlands

Abstract

Deep learning models have achieved promising results in breast cancer classification, yet their 'black-box' nature raises interpretability concerns. This research addresses the crucial need to gain insights into the decision-making process of convolutional neural networks (CNNs) for mammogram classification, specifically focusing on the underlying reasons for the CNN's predictions of breast cancer. For CNNs trained on the Mammographic Image Analysis Society (MIAS) dataset, we compared the post-hoc interpretability techniques LIME, Grad-CAM, and Kernel SHAP in terms of explanatory depth and computational efficiency. The results of this analysis indicate that Grad-CAM, in particular, provides comprehensive insights into the behavior of the CNN, revealing distinctive patterns in normal, benign, and malignant breast tissue. We discuss the implications of the current findings for the use of machine learning models and interpretation techniques in clinical practice.

Data and Code Availability This paper uses the publicly available Mammographic Image Analysis Society (MIAS) dataset (Suckling et al., 2015) which was acquired through the Kaggle platform (mias-mammography dataset). The MIAS dataset contains 322 images of full mammography scans, the corresponding labels, and region of interest (ROI) annotations. The dataset is licensed under CC BY 2.0 UK. A preprocessed version of the MIAS dataset that was used for the current analyses along with the relevant preprocessing code is available on GitHub (link).

Institutional review board (IRB) The research presented here did not require IRB approval.

1. Introduction

Breast cancer is the most prevalent cancer in women, and early detection through various methods, including mammography, is crucial to decrease mortality rates (Global Cancer Observatory, 2020; Mughal et al., 2018). Mammography involves taking X-ray images of breast tissue to identify abnormalities such as calcifications and masses (National Cancer Institute, 2024). These breast abnormalities can be benign, non-cancerous abnormal growths not spreading outside the breast or malignant, cancerous tumors that can spread to other organs. The accurate classification of benign or malignant breast lumps is crucial due to the health risks of malignancy, yet, it is even challenging for skilled radiologists. In fact, false negatives within mammography screening can be attributed, among other factors, to human perception errors (Ekpo et al., 2018) and misinterpretations (Palazzetti et al., 2016). This underscores the potential of computer-aided diagnosis (CAD) and deep learning systems in assisting radiologists to more accurately interpret mammographic images, thereby potentially mitigating these diagnostic errors (Lee et al., 2013; Coccia, 2020).

Deep learning (DL) algorithms can automatically extract features and effectively represent high-dimensional data. Convolutional neural networks (CNNs), types of DL models, are particularly suitable for image recognition tasks. They have shown promising results in classifying breast cancer from screening mammograms (El Houby and Yassin, 2021; Li et al., 2019). Enhancements in CNN performance involve various strategies, such as pretraining on image patches extracted from the regions of interest (ROI) in mammographic images containing breast abnormalities (Shen et al., 2019). This approach demonstrated improved model performance when subsequently training on full mammograms

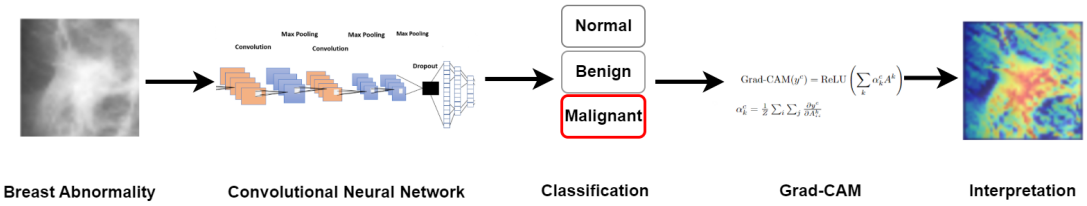


Figure 1: Proposed System: A CNN-based Mammography Classification Framework augmented with Grad-CAM for enhanced explainability.

without ROI annotations. Further, transfer learning across different mammographic datasets, as well as using pretrained models on ImageNet, such as InceptionV3 (Szegedy et al., 2016), has been shown to outperform models trained from scratch for detecting breast masses (Agarwal et al., 2019). To address overfitting concerns, data augmentation - which involves creating new training samples - has been employed to enhance model robustness (Li et al., 2019).

DL, while advancing the accuracy of diagnostic models, comes at the cost of interpretability, also known as the 'black box' problem. This issue arises as deep learning uses higher-level abstractions in deep hidden layers, which, although increasing the model accuracy, make the reasoning behind its decisions more opaque to human understanding. As a response to a lack of model interpretability, a new branch in artificial intelligence (AI) research has focused on explainable AI (XAI) which aims to provide insight into the model predictions to make deep learning models more interpretable.

Recent research has explored the interpretability of neural networks, using numerical datasets such as the Wisconsin dataset which provides features describing breast abnormalities extracted from mammograms (Karatza et al., 2021; Hakkoum et al., 2021). These studies used various methods to enhance model transparency, such as Shapley values (Karatza et al., 2021), feature importance, and LIME (Hakkoum et al., 2021). However, with mammography relying on image analysis, the interpretability of CNNs that take images rather than numerical features as input is paramount. For instance, Zhang et al. (2018) developed interpretable end-to-end CNNs that encode semantic information for each filter, while Chen et al. (2019) proposed ProTopNet, a model that highlights image parts that motivated the model prediction by comparing prototypical image parts of a class.

However, while recent studies have focused on improving model accuracy in diagnosing breast cancer using mammographic images, the crucial aspect of model explainability has been overlooked. Although interpretable breast cancer models have been applied to numerical breast cancer datasets, the post-hoc explainability of CNNs classifying mammographic images has, to the best of our knowledge, not been explored yet. As end-to-end intrinsically interpretable DL models might trade off accuracy, this underscores the importance of post-hoc interpretability of established image classifiers, such as CNNs, to enhance transparency in DL-driven breast cancer classification.

This research seeks to bridge the gap in explainable AI (XAI) within mammography, emphasizing the critical yet often neglected aspect of explainability in AI-driven diagnostics. By evaluating three post-how interpretability algorithms (Kernel SHAP, LIME, and Grad-CAM) following the training of a CNN on the MIAS mammogram dataset, our work not only addresses the automated classification of normal, benign, and malignant breast tissue, but integrates post-hoc explainable AI techniques to uncover the CNN's predictive rationale (Figure 1). We highlight Grad-CAM's capability to reveal distinct patterns between breast abnormalities, emphasizing the potential of explainable AI in fostering trust and transparency in CNN-based mammogram classification. Our model offers deeper insights for automated classification, aiding radiologists with a transparent tool to improve clinical decision making, potentially enhancing patient outcomes and reducing diagnostic errors.

2. Methods

2.1. Dataset

The MIAS (Mammographic Image Analysis Society) dataset (Suckling et al., 2015) contains images of mammography scans and three corresponding labels: normal (0), benign (1), and malignant (2). Additionally, it provides the x and y coordinates and radius

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

Figure 2: MIAS mammograms with ROI annotations.

of the abnormality for benign and malignant images (Figure 2). The images are in PGM format and have a size of 1024 x 1024 pixels. They contain 161 pairs of RGB images resulting in 322 images in total of which 207 are normal, 64 are benign and 51 are malignant.

2.2. Preprocessing

The mammographic images underwent a preprocessing procedure adapted from El Houby and Yassin (2021). This preprocessing of the raw images (Figure 3(*a*)) included noise artifact removal (Figure 3(*b*)) and image quality enhancement using contrast limited adaptive histogram equalization (CLAHE) (Zuiderveld (1994), (Figure 3(*c*)). For abnormal images, the region of interest (ROI) was extracted using the x and y coordinates and radius of the abnormality (Figure 3(*d*)), while normal images were cropped using a central breast area. We removed four abnormal images without ROI annotations from the dataset. After cropping, all images were resized to 224 x 224 pixels and visually inspected. Two ROI images with incorrect coordinate sets were identified and removed from the dataset, resulting in a final dataset of 316 images.

2.3. Splitting, data augmentation, balancing

We split the dataset into a training (proportion of images: 0.70), validation (0.15), and test set (0.15). To ensure a representative proportion of each class in every set, we applied a stratified split. To mitigate potential overfitting we augmented the dataset by creating new samples through small transformations of the original data (Perez and Wang, 2017; Oza et al., 2022; Li et al., 2019). This included rotation (0°, 90°, 180°, 270°) (Figure 3(*e*)), vertical flipping (Figure 3(*f*)) and random brightness and contrast changes in the ranges (-15,15) and (0.5,1.5) respectively (Figure 3(*g*)). After removing two duplicates, data augmentation resulted in a training set of 3534 images. The training dataset was subsequently balanced, yielding a dataset consisting of 1566 images for training (528 normal, 528 benign, 528 malignant images), 47 images (31 normal, 9 benign, 8 malignant) for validation, and 48 images for testing (31 normal, 9 benign, 8 malignant).

2.4. CNN architecture and hyperparameter tuning

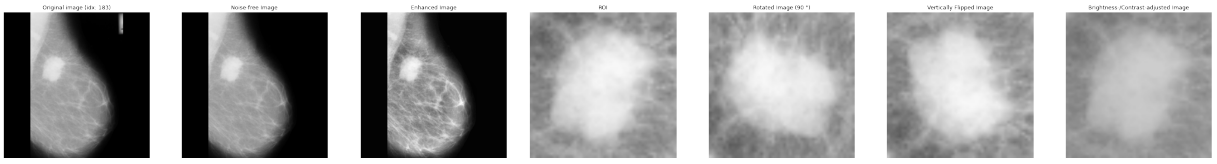
We fit a CNN to the data. CNNs use convolutional layers to extract features by sliding over the image and producing a feature map for features on different locations of the image. The CNN architecture used in this study was adapted from El Houby and Yassin (2021), who revealed promising results for this architecture on the MIAS dataset (accuracy: 0.95 for binary classification). The architecture of the CNN is visualized in Figure 4 and a detailed description of it is provided in Table A1. Prior to training, we normalised image pixels to [0,1] and one-hot encoded class labels. The learning rate and batch size were set at 0.0001 and 16, following Kandel and Castelli (2020). We used the Adam optimizer to optimize computational efficiency and minimize parameter tuning (Kingma and Ba, 2014). To improve the generalization performance of the model, a dropout layer with a rate of 0.5 was applied before the last fully connected layer (Li et al., 2018). Finally, we used the categorical cross-entropy loss function.

To improve model performance, hyperparameter tuning for the number of epochs and class weights was performed. We employed early stopping to balance the risk of overfitting, haltering training when the validation loss ceased to decrease for 10 successive epochs. The model initially trained for 50 epochs. The validation loss of the first CNN model (here referred to as 'CNN_1') stopped improving after 14 epochs (validation loss: 0.50; validation accuracy: 0.81), thus these weights were restored and used for the model predictions. Class weight tuning was critical due to the model's sensitivity to these parameters. Besides a default uniform class weight, other weight configurations prioritizing malignant cases (e.g., 0:1, 1:2, 2:3; 0:2, 1:3, 2:4) were tested and compared. Class weights of 0:1, 1:2, 2:3 were selected, assigning higher weights to benign and malignant cases.

3. Analysis

3.1. Model Evaluation Metrics

We evaluated the performance of the CNN using class-specific recall, precision, and F1-scores. In light



(a) Image (b) Noisefree (c) Enhanced (d) ROI (e) 90°rotated (f) Flipped (g) Brightness/ Contrast

Figure 3: Preprocessing procedure for all images, visualized for an example image (idx:183), depicting the original image (a), a preprocessed versions after noise removal (b), image enhancement (c), extractions of the Region of Interest (ROI) (d), 90° rotation (e), vertical flipping (f), and brightness and contrast changes (g).

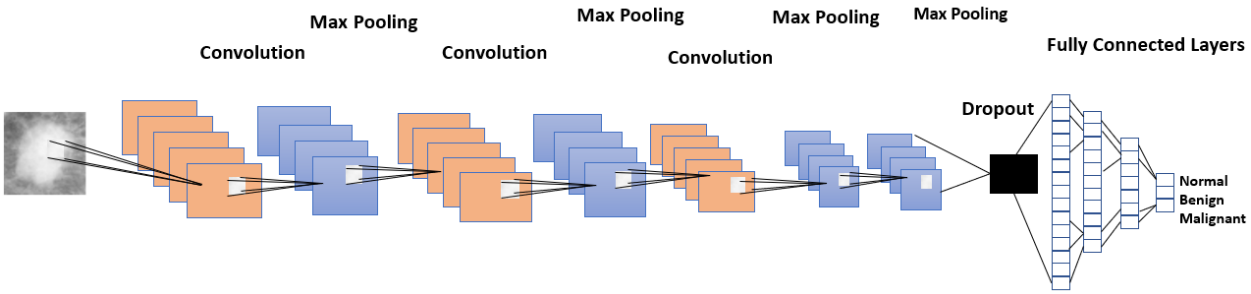


Figure 4: CNN architecture adapted from El Houby and Yassin (2021). ROIs are fed into a network consisting of three convolutional layers and four pooling layers, followed by a flattening layer with three fully connected layers. The output layer uses a softmax activation function to classify the ROIs as normal, benign, or malignant.

of the class imbalance in the validation and test sets, averaged macro metrics and class-specific metrics (accuracy, recall, precision, F1-score, balanced accuracy) were computed. To leverage the sensitivity of model evaluation metrics to the initialization of the CNN, average performance metrics were computed over ten iterations of training and evaluation and compared against a majority class baseline (Jaamour, 2020). Furthermore, the area under the receiver operating characteristics (AU-ROC) curve, which depicts the trade-off between true positive and false positive rates, was computed.

3.2. Model Interpretability

To investigate the interpretability of the CNN model, three model interpretability techniques were applied: Kernel SHAP, Grad-CAM, and LIME. Below, we describe each of these techniques in more detail.

3.2.1. SHapley Additive exPlanations (SHAP)

SHAP is a model interpretation technique introduced by Lundberg and Lee (2017) and inspired by Shapley values, a concept from cooperative game theory to calculate feature importance $\phi_i(f, x)$. It quantifies feature i's marginal contribution ϕ_i to the actual model's prediction for input value x. It does so by taking the difference in the model prediction before ($f(S)$) and after ($f(S \cup \{i\})$) feature i is added. Shapley values are weighted and averaged across all possible feature combinations. Formally, the feature importance $\phi_i(f, x)$ is defined as:

$$\phi_i(f, x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(F - |S| - 1)!}{F!} (f(S \cup \{i\}) - f(S)) \tag{1}$$

where F is the total number of features in the input space and S is the total number of feature coalitions before i.

We calculated SHAP values using Kernel SHAP, which is a computationally efficient, model-agnostic method to approximate SHAP values for any black-box model (including CNNs) using a linear regression model (Lundberg and Lee, 2017). We applied Kernel SHAP to segments of the mammographic images obtained via the SLIC (simple linear iterative clustering) algorithm (Achanta et al., 2012) using k-means

clustering with 1000 iterations. After evaluating various hyperparameter settings, we selected a configuration of 100 segments with a compactness value of 1, which optimally balances the trade-off between spatial and color proximity. This setup was particularly effective in capturing the edges and shapes of abnormalities in the mammographic images.

3.2.2. Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM is a gradient-based technique to provide visual explanations for a CNN model by highlighting image regions that are important for the model prediction (Selvaraju et al., 2017). The algorithm first computes the gradients of the output class score y^c with respect to the feature map activations A^k_{ij} of the k-th convolutional layer of the CNN. The gradients $\frac{\partial y^c}{\partial A^k_{ij}}$ describe how much each element of the activation map contributes to the prediction of class c. Averaging all gradients over all spatial locations (i, j) of the feature map, with a normalization factor Z for the total number of pixels in the activation map, yields the neuron importance weights α^c_k assigned to the k-th activation map for class c. To obtain the Grad-CAM localization map, the weights α^c_k are then used to weight the feature activation maps A^k, producing a weighted sum that is passed through a ReLU activation function:

$$\text{Grad-CAM}(y^c) = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right) \qquad (2)$$

where $\alpha^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}$

As recommended by Selvaraju et al. (2017), we used the last convolutional layer ($k = 3$) since it captures high-level features, represented in deeper convolutional layers, while preserving spatial information that is lost in the subsequent fully connected layer.

3.2.3. Local Interpretable Model-Agnostic Explanations (LIME)

LIME suggests local interpretable models that can approximate any complex 'black-box' classifier for a specific data point (Ribeiro et al., 2016). The method produces perturbated samples close to the original data point which are then converted to interpretable representations.

An explanation for the original data point x is sought given a model f with the predicted class $f(x)$. A local interpretable linear model g is used to approximate f, with the proximity measure π_{x_0} indicating the similarity between x and the perturbated instances z used to train g. LIME aims to minimize the loss $L(f, g, \pi_{x_0})$, which describes how well g approximates f in the neighborhood of x while restricting the model complexity $\Omega(g)$. The regularization term $\Omega(g)$ penalizes complex models. For linear models, for example, complexity is defined as the number of non-zero coefficients. The explanation produced by LIME at a local point x is formally defined as:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_{x_0}) + \Omega(g) \qquad (3)$$

The conceptual intuition between LIME is visualized in Figure 5. The bold star in Figure 5 represents the data point x for which an explanation is sought. Red stars are generated samples in the neighborhood of x belonging to the pink class, whereas blue circles describe perturbated samples that belong to the blue class. The size of the perturbated samples represents the proximity to x. The complex decision boundary of the model f is depicted as the border of the pink and blue background, whereas the black dashed line is the explainable model g that approximates f.

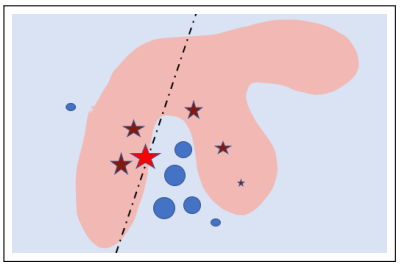


Figure 5: The intuition behind LIME. Adapted from Ribeiro et al. (2016).

3.3. Evaluation of Interpretability Algorithms

We evaluate the ability of LIME, Kernel SHAP, and Grad-CAM to explain the predictions of the CNN model through several criteria. These criteria are 1) computational efficiency, estimated on the basis of running time to determine time complexity under practical time constraints; 2) robustness, assessed by the consistency of explanations for a single input across multiple runs; 3) quality of visual explanations (masks and heatmaps) generated by each algorithm, established through the alignment of explanations with lesioned areas in mammograms.

4. Results

Below, we describe the results of the analysis. The result section is divided into two parts. First, we present the performance of the CNN. Second, we evaluate the three interpretability algorithms: Kernel SHAP, LIME, and Grad-CAM.

4.1. CNN model performance

As noted above, the CNN model was evaluated on 48 mammographic test images and for a more robust evaluation, the averaged results of multiple independent CNN runs are reported. The overall evaluation metrics for the performance of the CNN model are presented in Table 1. The CNN model outperforms a majority-predicting baseline model achieving an overall accuracy of 0.77. The negligible difference between macro recall (0.66) and macro precision (0.64) suggests a good balance between capturing true positive cases (recall) and correctly classifying positive predictions (precision).

Class-specific performance is presented in Table 2. The model achieved the best results for normal cases (Class 0; F1 = 0.93, AUC: 0.98), while the model's performance was least good for malignant cases (Class 2; F1 = 0.44, AUC: 0.83). The benign class was predicted moderately well (Class 1; F1 = 0.58, AUC: 0.90). The class-specific AUC curves are visualized in Figure A1 in Appendix B. All class-specific evaluation metrics exceeded baseline performance, indicating that the model learned discriminative patterns across the different classes to at least some extent. An overview of true classes versus predicted classes is presented in the confusion matrix in Figure 6. Overall, the average number of correct predictions was 37.2, whereas the average number of incorrect predictions was 10.8.

4.2. Performance of evaluation algorithms

Below, we present the evaluation of the performance of the interpretability algorithms Kernel SHAP, LIME, and Grad-CAM in terms of computational efficiency, robustness, and quality of visual explanations.

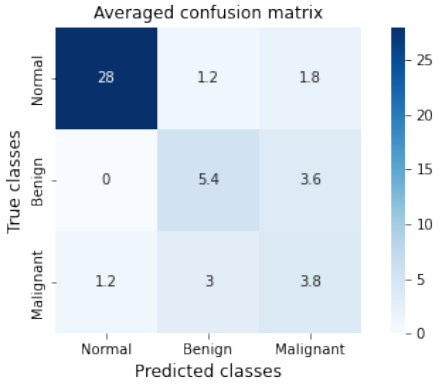


Figure 6: Averaged confusion matrix.

4.2.1. Computational efficiency

The efficiency of each interpretability algorithm was evaluated based on the average running time per image, as presented in Table 3. Running times between algorithms differed substantially. Grad-CAM was the fastest algorithm, with an average running time of 0.29 seconds per image. By contrast, LIME required an average of 50.5 seconds per image to compute an explanation.

4.2.2. Robustness

Whereas Grad-CAM is a deterministic algorithm, LIME and SHAP are sampling-based techniques. As a consequence, results for an image vary between different runs of both of these algorithms. The robustness analysis revealed that differences between runs are substantial for LIME. An example of this is presented in Figure 7(*a*), which visualizes the explanations computed by LIME for five runs on the same image. As can be seen in Figure 7(*b*), the differences across runs were more subtle for SHAP.

4.2.3. Comparison of visual explanations

The effectiveness of LIME, Kernel SHAP, and Grad-CAM was examined for a correctly classified benign mammogram, with typical oval shape and smooth borders, and for a malignant abnormality, with char-

Table 1: Average overall evaluation metrics

	Macro Precision	Macro Recall	Macro F1-score	Overall Accuracy	Balanced Accuracy
Majority baseline	0.00	0.33	0.00	0.65	0.33
CNN model	0.64	0.66	0.65	0.77	0.66

Table 2: Averaged per-class evaluation metrics

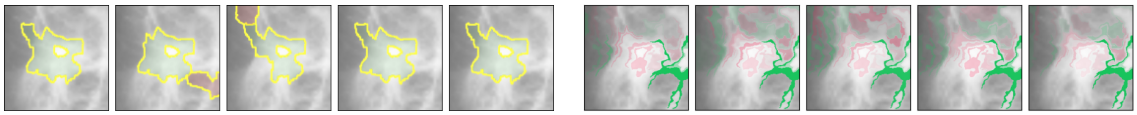
	Class 0	Class 1	Class 2
Precision	0.96	0.56	0.41
Recall	0.90	0.60	0.48
F1-score	0.93	0.58	0.44
Baseline F1-score	0.79	0.00	0.00

Table 3: Average running time (s) per image for LIME, Grad-CAM, and Kernel SHAP

LIME	Grad-CAM	Kernel SHAP
50.5	0.29	4.48

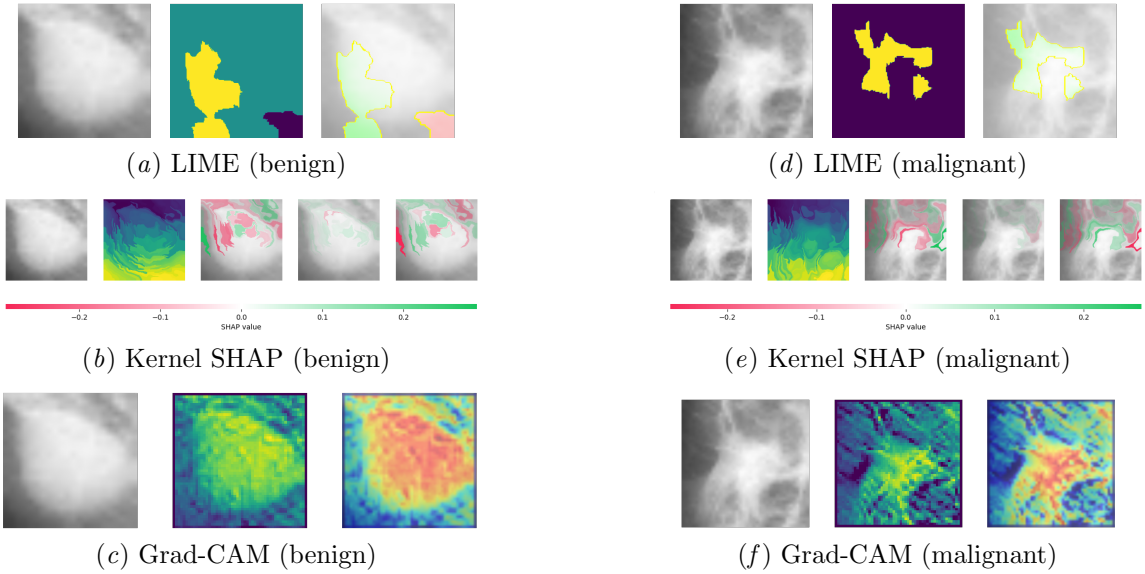
acteristic irregular, tentacle-like patterns. The corresponding explanations are visualized in Figure 8.

LIME, highlighting positively and negatively contributing areas to the prediction in green and red, revealed partial coverage with the lesions only. Kernel SHAP offers separate explanations for benign (third image in panel), malignant (fourth image in panel), and normal (fifth image in panel) predictions - using green and red to signify superpixels that increase or decrease class probabilities. As can be seen in Figure 8, it succeeded in identifying key edges in both benign (Figure 8(b)) and malignant abnormalities (Figure 8(e)) that overlap with the lesion borders. Grad-CAM, employing a red-to-blue heatmap, highlights pixels that were the most and least influential in the classification process. The explanations of Grad-CAM revealed significant overlap with the actual shapes of both benign (Figure 8(c)) and malignant abnormalities (Figure 8(f)) and were closely aligned with the characteristics of the lesion. Appendix C presents the explanations for LIME, Kernel SHAP, and Grad-CAM for an incorrect prediction (i.e., a malignant image classified as benign).



(a) LIME (b) Kernel SHAP

Figure 7: LIME (a) and Kernel SHAP (b) both produce different results over multiple runs.

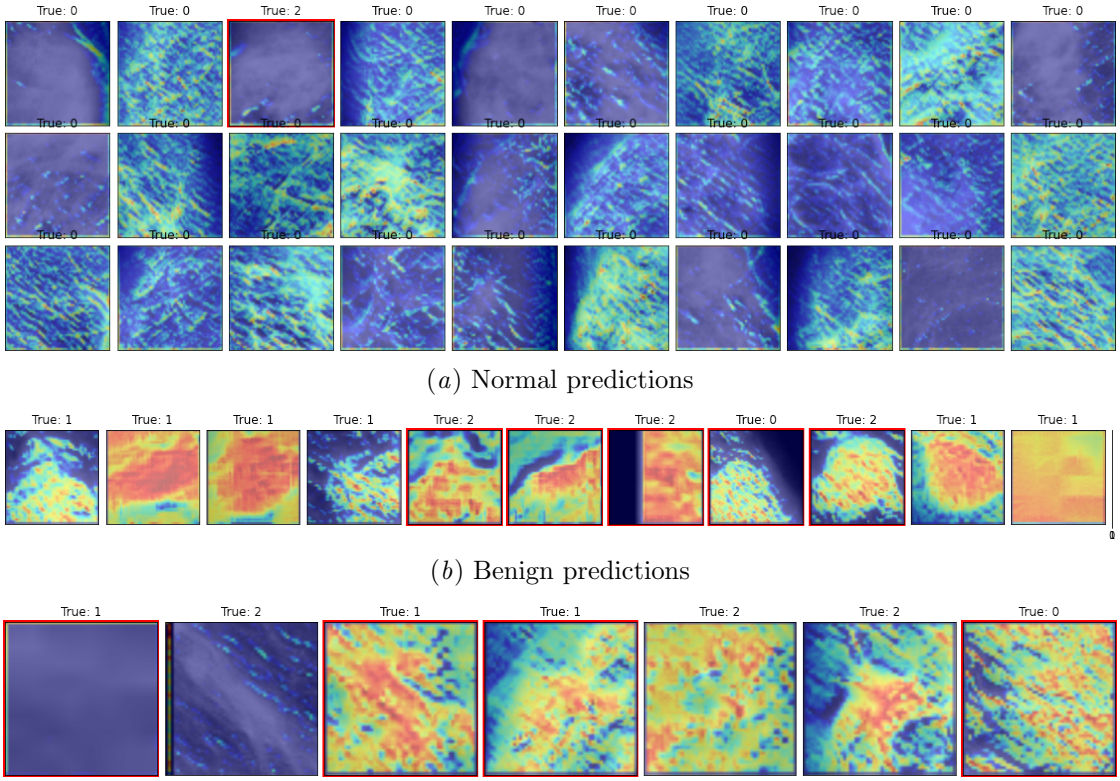


(a) LIME (benign) (d) LIME (malignant)

(b) Kernel SHAP (benign) (e) Kernel SHAP (malignant)

(c) Grad-CAM (benign) (f) Grad-CAM (malignant)

Figure 8: Original image (left of panel), heatmap (middle of panel) and overlaid heatmap (right of panel) for the explanations of the LIME (a, d), Kernel SHAP (b, e), and Grad-CAM (c, f) algorithms for a benign (left) and malignant (right) mammogram.



(*a*) Normal predictions

(*b*) Benign predictions

(*c*) Malignant predictions

Figure 9: Grad-CAM predictions. Red boxes indicate incorrect predictions.

Given its excellent performance, we present the explanations offered by the Grad-CAM algorithm in more detail in Figure 9, which shows Grad-CAM explanations for all mammographic images in the dataset, categorized by predicted class. Images for which the predictions of the CNN were incorrect are marked by a red border. Analogous figures for LIME (Figure A3) and Kernel SHAP (Figure A4) are presented in Appendix D. Explanations for normal predictions are characterized by a lack of distinct patterns (Figure 9(*a*)), explanations for true benign cases by round or oval shapes (Figure 9(*b*)), and malignant cases - with two notable exceptions - by more irregular, undefined shapes (Figure 9(*c*)). As such, Grad-CAM is able to differentiate between the three classes of images effectively.

5. Discussion

We presented a study of the model interpretation algorithms LIME, Grad-CAM, and Kernel SHAP in the context of a CNN fit to the MIAS dataset for mammographic image classification. Below, we discuss the findings of this study in more detail. The discussion is divided into four sections: the evaluation of the CNN, the evaluation of the explanations offered by the three interpretation algorithms that help gain insight into the CNN's 'black-box' model predictions, practical implications for clinical relevance and directions for future research.

5.1. Evaluation of the CNN

We applied a variety of preprocessing techniques to enhance the MIAS mammogram dataset's usability for CNN classification, including noise removal, image contrast enhancement, and ROI extraction, supplemented by data augmentation addressing the original dataset size limitations. To reduce model bias towards the majority class, the classes were balanced. This preprocessed dataset is now publicly available, promoting research transparency and alleviating time-intensive preprocessing requirements.

However, the dataset's inherent limitations, including low image quality and the absence of a second breast view potentially negatively affected the performance of the CNN. The reliance on manual ROI annotations and classification labels, while necessary, also introduces potential biases, as mammogram interpretation is subject to human error, especially for rare cancer types that have been shown to remain undetected by radiologists (Evans et al., 2013). Future research could explore the fine-tuning of the trained ROI model on whole mammogram images as suggested by Shen et al. (2019), to enable the classification of complete mammogram images in the absence of ROI annotations. Given sufficient computational resources, mammograms that offer higher resolution and two breast views such as the INBreast (Moreira et al., 2012) or DDSM dataset, (Heath et al., 1998), could therefore provide finer-grained information to improve the classification performance (Petrini et al., 2022).

The CNN effectively distinguished between normal, benign, and malignant cases with few misclassified normal cases (recall = 0.90). However, it showed a higher prevalence of misclassified malignant lumps (recall = 0.48), leaving more than half of true cancer cases undetected. This highlights the need for improved malignancy detection - as false negatives can have severe negative implications such as treatment delays or a false sense of security as the women are unaware of their disease (National Cancer Institute, 2023). As the majority of incorrect classifications involved the distinction between benign and malignant cases, future studies should consider a more fine-grained classification between these two classes, such as a five-class classification, distinguishing calcifications from masses (Shen et al., 2019). Additionally, deploying deeper CNNs with smaller kernel sizes (Li et al., 2019) and exploring more advanced architectures such as Vision Transformers (Dosovitskiy et al., 2020) could further reduce false positives.

Our use of an averaged confusion matrix mitigated the influence of variations across different training runs, thus enhancing metric reliability by capturing a more general trend. Yet, training and evaluating the model on a single data split may have biased the results towards that specific split. While our findings suggest that the CNN identifies characteristic tumor shapes, we acknowledge our dataset's limited diversity and stress the importance of further evaluations on broader datasets. Such assessments are crucial to verify the CNN's capability to accurately distinguish between various breast cancer types, including rare and complex cases, across different patient populations and screening conditions. Furthermore, the approach as suggested by Agarwal et al. (2019) could be adapted to not only test but fine-tune the model by training on multiple mammographic datasets, potentially improving its performance and making the model more robust to nuances present in different datasets.

5.2. Evaluation of explanations

The main objective of the current study was to provide better insight into the reasons why the CNN classified a mammogram as normal, benign, or malignant. Specifically, we compared the algorithms Grad-CAM, SHAP, and LIME with respect to their computational efficiency, robustness, and quality of explanations. When adding an interpretation algorithm to a breast cancer detection procedure it is of vital importance that the algorithm offers explanations that are computationally efficient as well as consistent. Time-efficient explanations are crucial because radiologists face time constraints during clinical decision-making. Additionally, trust into an AI-powered technology is an important factor for clinicians to adopt it in practice (Tucci et al., 2022). Consistency of an interpretation algorithm, therefore, is essential (Yu, 2013; Lee et al., 2019).

In light of these considerations, LIME - which takes an average of 50.5 seconds to compute an explanation and generates highly unstable masks - may be less suitable in clinical practice. SHAP, similar to LIME, is a perturbation-based approach that involves the generation of new data points as part of the explanation process and thus renders explanations in a non-deterministic manner. Nonetheless, SHAP explanations revealed less variability as compared to LIME explanations. Furthermore, SHAP was more computationally efficient, as indicated by reduced running times. Optimal performance in terms of computational efficiency and robustness, however, was observed for Grad-CAM, which takes less than one second (0.29 seconds) per image to generate explanations and provides deterministic explanations by directly utilizing the model gradients. Grad-CAM thus appears to be highly suitable for practical implementation.

Benign and malignant cancer appears as white or light-grey matter on mammograms and thus sets itself apart from the grey breast tissue. Based on

this understanding, the assumption was made that benign and malignant abnormalities are characterized by white lumps on mammograms, disregarding the background. LIME and Kernel SHAP were unable to accurately identify the entire breast lesion. By contrast, Grad-CAM identified lesions (most) accurately and generated (the most) plausible explanations by effectively identifying distinctive features corresponding to normal, benign, and malignant breast tissue characteristics. The Grad-CAM heatmaps for benign predictions showed clearly defined round/oval shapes, indicating the presence of non-spreading, non-cancerous abnormal tissue. The heatmaps for malignant predictions displayed less-defined, scattered areas spreading into different directions, aligning with the invasive characteristic of cancerous malignant cells into the surrounding areas. Normal predictions did not exhibit distinctive patterns, consistent with the expectation that healthy breast tissue does not exhibit suspicious abnormalities. As such, the explanations offered by Grad-CAM fit well with the patterns observed in the diagnosis of breast cancer abnormalities in clinical practice (Global Cancer Observatory, 2020).

5.3. Practical Implications for Clinical Relevance

The suitability of Grad-CAM in providing explanations aligning with human intuitions have practical implications for the clinical practice. We propose a system augmenting a CNN-based classifier with visual explanations in the form of heatmaps, serving as a foundational tool for future research by incorporating interpretability within automated breast cancer classification. This does not only enhance transparency of DL-based automated mammography diagnosis, but can furthermore serve as an educational tool for medical trainees by providing large number of examples of characteristic heatmaps of breast cancer types. Additionally, our framework could be extended to not only identify well-known features of known breast abnormalities but also uncover novel diagnostic markers of rare cancer types.

5.4. Future Directions

We presented an initial exploration of interpretation algorithms in the context of mammography image classification. In this initial exploration, we focused on the qualitative evaluation of different interpretability algorithms, acknowledging the possibility of observer bias. Future research, inspired by previous experiments conducted by Ribeiro et al. (2016) and Selvaraju et al. (2017) could and should involve human-subject experiments to more reliably assess the extent to which Grad-CAM enhances trust of radiologists in deep learning-based technologies. Additionally, further investigation is in order to evaluate to what extent Grad-CAM is effective in detecting human diagnostic errors, particularly in identifying misclassified benign or malignant cases and in the identification of rare cancer types that may be missed by radiologists (Evans et al., 2013). Furthermore, it is important to note that we only compared post-hoc methods in the current study. Post-hoc methods are approximations of the original model and don't modify the 'black-box' architecture of the CNN model itself. The use of 'white-box' models, intrinsically interpretable CNNs (see, e.g. Zhang et al., 2018), is another avenue to explore in future research.

Finally, evaluating interpretability techniques still relies on the expertise and judgment of the observer, as there are no standardized qualitative or quantitative measures to assess the plausibility of provided explanations. For future research, we therefore recommend the involvement of domain experts, such as professional radiologists, to establish a ground truth. More specifically, since the ROIs in the MIAS dataset lack exact lesion masks, professional radiologists could mark the precise breast abnormalities present in the ROI for a quantitative analysis of the predictions. This would enable the calculation of overlap between the interpretation mask and annotations provided by expert radiologists, using metrics like intersection over union (IoU) to measure the agreement between both.

6. Conclusions

We presented the application of three interpretation techniques - LIME, Kernel SHAPE, and Grad-CAM - to the results of a CNN trained on mammographic image data. Grad-CAM emerged as the preferred interpretation technique for the current data set and machine learning model, providing insightful explanations of the model's predictions in a time-efficient and stable way. The current findings provide valuable insights into the interpretability of deep learning models in the context of breast cancer classification and demonstrate the potential of explainable artificial intelligence (XAI) as a supplementary tool for radiologists in the diagnosis of breast cancer.

References

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11): 2274–2282, 2012.

Richa Agarwal, Oliver Diaz, Xavier Lladó, Moi Hoon Yap, and Robert Martí. Automatic mass detection in mammograms using deep convolutional neural networks. *Journal of Medical Imaging*, 6(3): 031409–031409, 2019.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

Mario Coccia. Deep learning technology for improving cancer care in society: New directions in cancer imaging driven by artificial intelligence. *Technology in Society*, 60:101198, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Ernest Usang Ekpo, Maram Alakhras, and Patrick Brennan. Errors in mammography cannot be solved through technology alone. *Asian Pacific journal of cancer prevention: APJCP*, 19(2):291, 2018.

Enas MF El Houby and Nisreen IR Yassin. Malignant and nonmalignant classification of breast lesions in mammograms using convolutional neural networks. *Biomedical Signal Processing and Control*, 70:102954, 2021.

Karla K Evans, Robyn L Birdwell, and Jeremy M Wolfe. If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening. *PloS one*, 8(5):e64366, 2013.

Global Cancer Observatory. Cancer today. International Agency for Research on Cancer - World Health Organization, 2020. URL https://gco.iarc.fr/today/home. Accessed: 2023-04-15.

Hajar Hakkoum, Ali Idri, and Ibtissam Abnane. Assessing and comparing interpretability techniques for artificial neural networks breast cancer classification. *Computer methods in biomechanics and biomedical engineering: imaging & visualization*, 9 (6):587–599, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Michael Heath, Kevin Bowyer, Daniel Kopans, Philip Kegelmeyer, Richard Moore, Kyong Chang, and S Munishkumaran. Current status of the digital database for screening mammography. *Digital Mammography: Nijmegen, 1998*, pages 457–460, 1998.

Adam Jaamour. Breast cancer detection in mammograms using deep learning techniques [master's thesis, university of st andrews, school of computer science]. 2020. URL https://info.cs.st-andrews.ac.uk/student-handbook/files/project-library/cs5098/agj6-Final_report.pdf.

Ibrahem Kandel and Mauro Castelli. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT express*, 6(4):312–315, 2020.

Panagiota Karatza, Kalliopi Dalakleidi, Maria Athanasiou, and Konstantina S Nikita. Interpretability methods of machine learning algorithms with applications in breast cancer diagnosis. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2310–2313. IEEE, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Cindy S Lee, Paul G Nagy, Sallie J Weaver, and David E Newman-Toker. Cognitive and system factors contributing to diagnostic errors in radiology. *American Journal of Roentgenology*, 201(3): 611–617, 2013.

Eunjin Lee, David Braines, Mitchell Stiffler, Adam Hudler, and Daniel Harborne. Developing the sen-

sitivity of lime for better machine learning explanation. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, pages 349–356. SPIE, 2019.

Bin Li, Yunhao Ge, Yanzheng Zhao, Enguang Guan, and Weixin Yan. Benign and malignant mammographic image classification based on convolutional neural networks. In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, pages 247–251, 2018.

Hua Li, Shasha Zhuang, Deng-ao Li, Jumin Zhao, and Yanyun Ma. Benign and malignant classification of mammogram images based on deep learning. *Biomedical Signal Processing and Control*, 51: 347–354, 2019.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.

Bushra Mughal, Muhammad Sharif, Nazeer Muhammad, and Tanzila Saba. A novel classification scheme to decline the mortality rate among women due to breast tumor. *Microscopy research and technique*, 81(2):171–180, 2018.

National Cancer Institute. Mammograms fact sheet. National Institutes of Health, 2023. URL https://www.cancer.gov/types/breast/mammograms-fact-sheet. Accessed: 2023-04-21.

National Cancer Institute. Breast cancer screening (pdq) - patient version. National Institutes of Health, 2024. URL https://www.cancer.gov/types/breast/hp/breast-screening-pdq. Accessed: 2024-01-21.

Parita Oza, Paawan Sharma, Samir Patel, Festus Adedoyin, and Alessandro Bruno. Image augmentation techniques for mammogram analysis. *Journal of Imaging*, 8(5):141, 2022.

Valentina Palazzetti, F Guidi, L Ottaviani, Gianluca Valeri, Silvia Baldassarre, and Gian Marco Giuseppetti. Analysis of mammographic diagnostic errors in breast clinic. *La radiologia medica*, 121:828–833, 2016.

Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

Daniel GP Petrini, Carlos Shimizu, Rosimeire A Roela, Gabriel Vansuita Valente, Maria Aparecida Azevedo Koike Folgueira, and Hae Yong Kim. Breast cancer diagnosis in two-view mammography using end-to-end trained efficientnet-based convolutional network. *Ieee Access*, 10:77723–77731, 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9 (1):12495, 2019.

John Suckling, J Parker, D Dance, S Astley, I Hutt, C Boggis, I Ricketts, E Stamatakis, N Cerneaz, S Kok, et al. Mammographic image analysis society (mias) database v1. 21. 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Victoria Tucci, Joan Saary, and Thomas E Doyle. Factors influencing trust in medical artificial intelligence for healthcare professionals: A narrative review. *J. Med. Artif. Intell*, 5(4), 2022.

Bin Yu. Stability. *Bernoulli*, 19(4), 9 2013. doi: 10.3150/13-bejsp14. URL https://doi.org/10.3150/13-bejsp14.

Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8827–8836, 2018.

Karel Zuiderveld. Contrast limited adaptive histogram equalization. *Graphics gems*, pages 474–485, 1994.

Appendix A. CNN architecture

Table A1: CNN Architecture adapted from El Houby and Yassin (2021). A dropout layer was added after the third fully-connected layer for model regularization purposes and the originally binary problem was transformed into a 3-class classification problem.

Layer #	Kernels	Kernel size	Stride	Padding	Output Shape	Output Size	# Parameters	Activation
Input Image	-	-	-	-	224,224,3	-	-	-
convo_1	16	5×5	1×1	0×0	220,220,16	665,856	1216	ReLU
max_pooling_1	-	2×2	2×2	-	110,110,16	166,464	0	-
convo_2	16	5×5	1×1	0×0	106,106,16	153,664	6416	ReLU
max_pooling_2	-	2×2	2×2	-	53,53,16	38,416	0	-
convo_3	14	3×3	1×1	1×1	53,53,14	33,614	2030	ReLU
max_pooling_3	-	2×2	2×2	-	26,26,14	8064	0	-
max_pooling_4	-	2×2	2×2	-	13,13,14	2016	0	-
flatten_1	-	-	-	-	-	2016	0	-
dense_1	-	-	-	-	512,1	512	1,211,904	-
dense_2	-	-	-	-	256,1	256	131,328	-
dense_3	-	-	-	-	128,1	128	32,896	-
dropout_1	-	-	-	-	512,1	-	-	-
dense_4	-	-	-	-	3,1	3	387	-
Total params:	-	-	-	-	-	-	1,386,177	-

The model architecture consists of three convolutional layers, each followed by max-pooling layers (two max-pooling layers after the third convolutional layer), with strides of 2 for pooling and 1 for convolution. Kernel sizes are 5x5 for the first two layers and 3x3 for the last, with 16, 16, and 14 kernels, respectively. The network incorporates ReLU activation functions, He weight initialization (He et al., 2015), three fully connected layers with 512, 256, and 128 units after the flattening layer, followed by a final softmax function for three-class (normal, benign, malignant) classification.

Appendix B. ROC curve

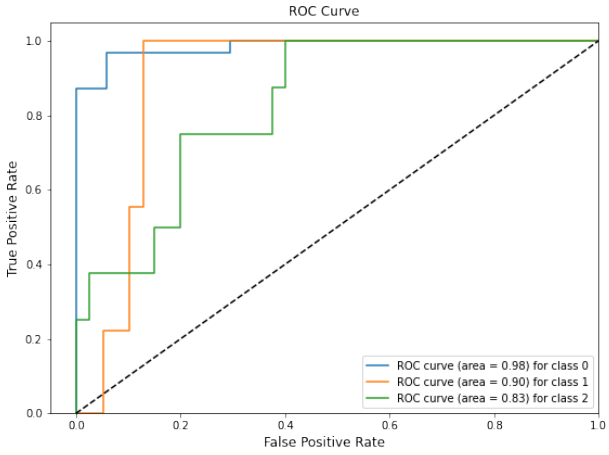
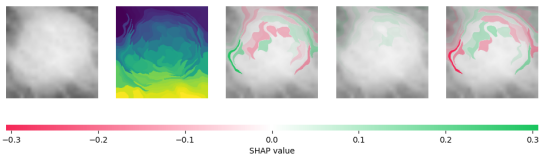


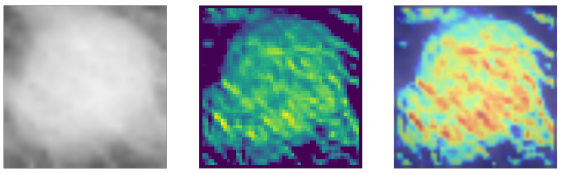
Figure A1: ROC curve for normal (blue), benign (orange), and malignant (green) predictions.

Appendix C. Explanations for an incorrect prediction

(a) LIME



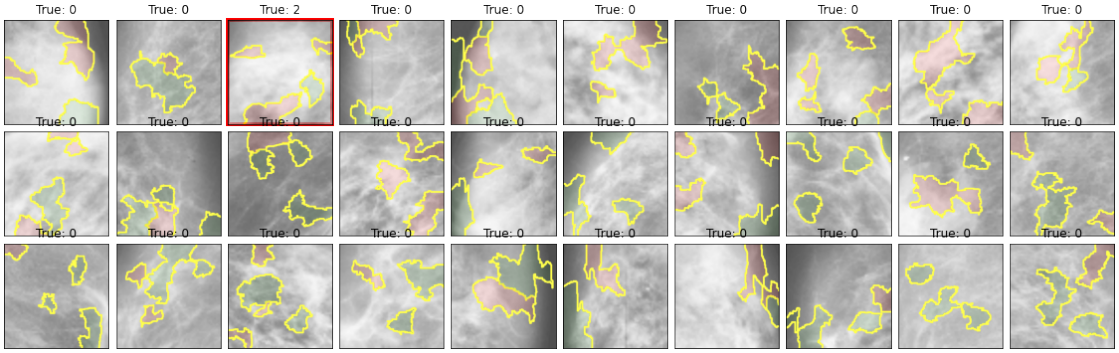
(b) Kernel SHAP



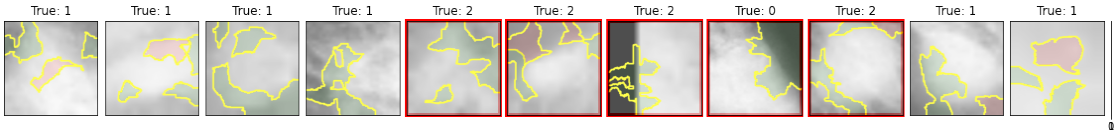
(c) Grad-CAM

Figure A2: Original image (left of panel), heatmap (middle of panel) and overlaid heatmap (right of panel) for the explanations of the LIME (a), Kernel SHAP (b), and Grad-CAM (c) algorithms for a malignant mammograms that was incorrectly predicted to be benign.

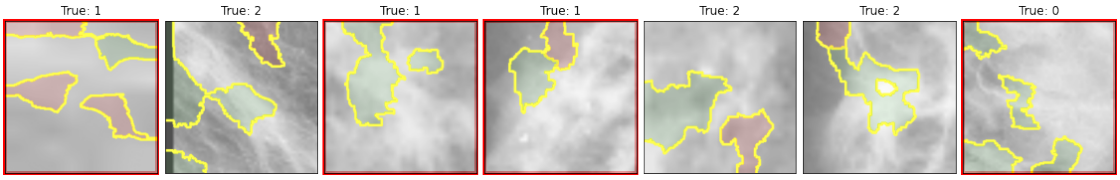
Appendix D. Predictions LIME and Kernel SHAP



(*a*) Normal predictions

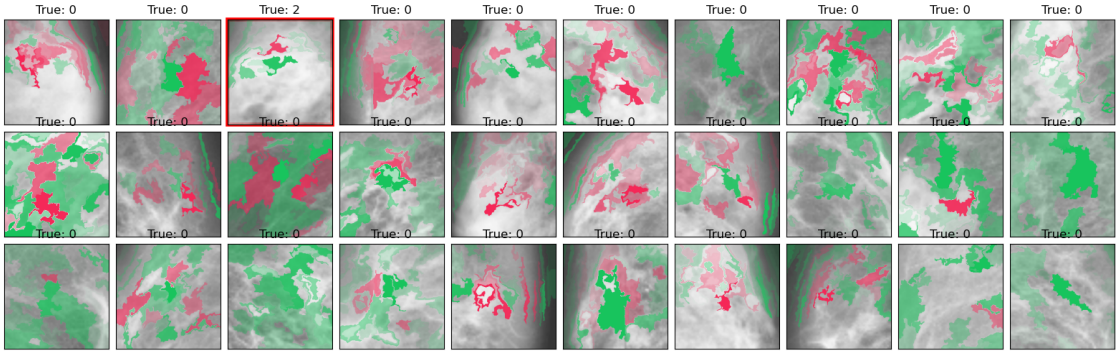


(*b*) Benign predictions

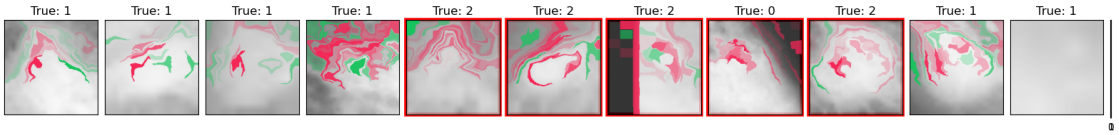


(*c*) Malignant predictions

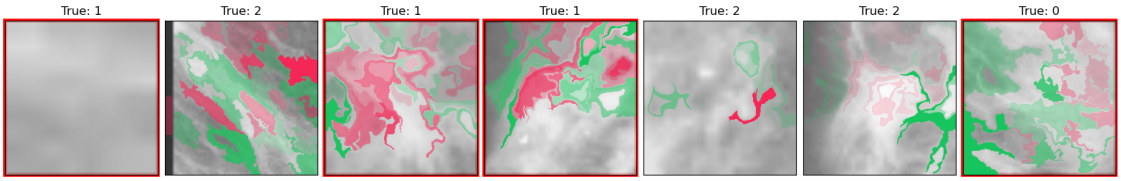
Figure A3: LIME predictions. Red boxes indicate incorrect predictions.



(*a*) Normal predictions



(*b*) Benign predictions



(*c*) Malignant predictions

Figure A4: Kernel SHAP predictions. Red boxes indicate incorrect predictions.