Evaluation of explainability tools and methods in medical diagnosis

Asfa Jamil Daniele Marini Luca Reggiani

jamil.asfa@ucy.ac.cy , {daniele.marini3 , luca.reggiani6}@studio.unibo.it

Abstract

Medical diagnosis plays a crucial role in patient care, but the lack of interpretability in complex machine learning models hinders their adoption in clinical settings. This paper evaluates various explainability tools and methods in the context of medical diagnosis. Firstly, we explore Model Agnostic techniques, such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations), which provide interpretability by approximating model behavior. Then, we report on three articles that discuss Explainability techniques specifically applied to breast cancer prediction and diagnosis. These articles shed light on the potential benefits and limitations of different approaches in this specific medical domain. Then we introduced backpropagation-based techniques for visual explanation, specifically Grad-CAM, some correlated approaches (CAM, Score-CAM) and lastly DeepLIFT. Finally we conducted experiments using Grad-CAM technique. The findings of this paper contribute to the evaluation and comparison of explainability tools and methods, providing insights into their potential application in medical diagnosis and opening avenues for further research in this critical domain.

1 Introduction

The advent of Convolutional Neural Networks (CNNs) and other deep learning models has facilitated significant advancements in different areas of computer vision, from image classification to object detection, semantic segmentation and visual question answering (VQA) [2].
In the recent years, increasing attention has been drawn to the internal mechanisms of CNNs, and the reason why the network makes specific decisions. This type of informations are essentials for building trust in intelligent systems and move towards their meaningful integration into our everyday lives.
The goal is to build "transparent" models that explain why they predict what they predict. This transparency is extremely useful for three reasons:

1. enhance the trust and the confidence of the users

2. allow us to identify issues of the model (and eventually fix them) or biases in the data

3. teaching a human about how to make better decisions (machine teaching) when the network is significantly stronger than humans (e.g. chess)

In certain fields, such as healthcare, it is essential to increase user trust and confidence, which is the aim of the techniques presented below.

We began by examining model-agnostic techniques, LIME [14] and SHAP [13], which approximate the behavior of black-box models to provide explanations.

Additionally, we report on three papers that discuss explainability techniques in breast cancer prediction and diagnosis. These papers include "Case-based Ensemble Learning System for Explainable Breast Cancer Recurrence Prediction"[8] "Explainable Artificial Intelligence for Breast Cancer: A Visual Case-Based Reasoning Approach"[11] and "Explain individual classification decisions"[3] Each paper explores different approaches to improve explainability in this medical domain.

Furthermore, we introduce backpropagation-based techniques, including Grad-CAM [17] and DeepLIFT [18]. These techniques utilize backpropagation to generate visual explanations, highlighting important regions in an image that influenced the model's decision-making process.

In order to gain a comprehensive understanding of Grad-CAM and its underlying principles, we also explored the concept of Class Activation Mapping (CAM) [21], from which Grad-CAM draws inspiration. Building upon this foundation, we further delved into Score-CAM [20], an innovative approach that addresses certain limitations present in Grad-CAM.

Lastly, in order to evaluate the effectiveness of backpropagation-based techniques, we specifically conducted experiments focusing on Grad-CAM. These experiments aim to assess its ability to provide visual explanations in the context of brain cancer detection.

2 Model Agnostic techniques

Model-agnostic approaches [15] are ways for creating explanations of the output of any machine learning model (aka black-box model), independent of the specific type of model used.

These techniques are designed to help users understand, trust and explain the decisions made by a model, regardless of its complexity or the nature of the data it was trained on.

Model-agnostic techniques are crucial because they provide increased transparency and explainability in AI systems, which is critical in many fields, including medical care and diagnosis.

Two well known approaches that are part of the model agnostic techniques are the LIME and the SHAP techniques.

- A prominent approach for creating local explanations of AI models is Local Interpretable Model-agnostic Explanations (LIME). LIME strives to give insight into a black-box model's decision-making process by developing a smaller, more easily understandable model that emulates the original model's behavior in a particular area of the feature space. LIME has been used to predict health outcomes and diagnose diseases in a range of medical contexts.

- SHapley Additive exPlanations (SHAP). SHAP is a framework for applying game theory to understand the result of whatever machine learning model. It gives a global interpretation of the model by giving relevance scores to each feature depending on how much it contributes to the output of the model.

2.1 LIME

In decision-making scenarios, it's crucial to have confidence in individual predictions and the model as a whole before deployment. Traditional accuracy metrics may not accurately reflect real-world performance. To address this, model agnostic techniques in XAI can provide explanations for individual predictions and overall model performance, offering qualitative insight into the connection between input features and the model's output. These techniques, such as feature importance analysis and surrogate modeling, can be applied to any machine learning model, making them widely applicable across various fields.

LIME [14] generates explanations for black-box models by perturbing input features to identify important features in the model's decision-making process. Perturbed instances are similar to the original instance but with variations in feature values. LIME evaluates each instance using a black-box classifier to obtain feature importance scores. A subset of the most important features is selected to train an interpretable model that can approximate the behavior of the black-box model in the local vicinity of a specific input instance. The interpretable model generates an explanation, highlighting important features and providing insights into the black-box model's behavior. The explanation is a summary of the important features and decision-making process of the black-box model for that instance. The process varies depending on the input data type, for example, LIME generates perturbed instances of images by adding random noise or cropping/rotating the image.

Mathematically, local surrogate models with interpretability constraint can be expressed as follows:

$$\text{explanation}(x) = \arg\min_{g \in G} L\left(f, g, \pi_x\right) + \Omega(g)$$

The explanation model for the instance x is the model g (the explainable one, e.g a linear regression model) that minimizes loss L (e.g. mean squared error), which measures how close the explanation is to the prediction of the original model f (a black-box model), while the model complexity $\Omega(g)$ is kept low (e.g. prefer fewer features). G is the family of possible explanations, for example, all possible linear regression models. The proximity measure π_x defines how large the neighbourhood around instance x is that we consider for the explanation. In practice, LIME only optimizes the loss part and the user has to determine the complexity by selecting the maxi-

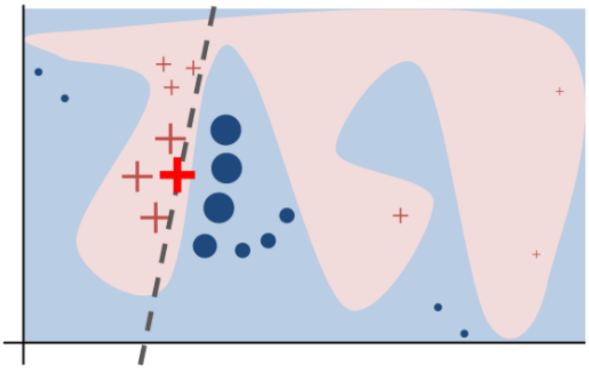


Figure 1: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

mum number of features that the explainable model may use.

In fact, very important in LIME is the number of the perturbed instances. Indeed, generating more perturbed instances can lead to better explanations, as it provides more information about the decision boundary of the black-box model, but it could be computationally expensive and may not provide significant additional information beyond a certain point.

In medical field, LIME is quite useful to understand the outcomes of the black-box models as well. In detail, in healthcare is widespread the use of Artificial Intelligence on images, for issues like cancer detection. So, it's important to understand how Lime works on images [7]. At first, LIME selects a region of interest (ROI) in the image. This can be done using various techniques, such as a superpixel algorithm [1]. LIME then generates perturbations of the image by modifying the selected region (with a different color or texture), and obtains predictions from the black-box model for each perturbed image. The difference between the predictions for the original image and the perturbed images is used to calculate the importance of each pixel in the selected region. This importance score is then used to fit a linear model, which approximates the behavior of the black-box model within the selected region. The linear model is interpretable, which means it can be used to explain how the black-box model makes its predictions for that particular image. The linear model is trained to predict the same output as the black-box model but only within the region of the image that LIME has selected. This means that the linear model is a simplified version of the black-box model, which only takes into account the most important features (pixels) within the selected region.

2.2 SHAP

SHAP stands for "SHapley Additive exPlanations" and is a model-agnostic approach to explain the output of any machine learning model. The SHAP approach is based on the concept of Shapley values from cooperative game theory, which assigns a value to each feature in a model based on how much it contributes to the prediction of the model. The SHAP approach works by first computing the Shapley values for each feature in the model, which quantifies the impact of each feature on the output of the model. These values are then used to provide explanations for individual predictions, such as by highlighting the top features that contributed most to the output.

The Shapley value of feature i, denoted as Φ_i, is the average marginal contribution of feature i over all possible coalitions of features that can be formed from the set of all features except i. This can be expressed as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

Where F is the set of all features except the feature i, S is a coalition (set) of features that does not include j, $|S|$ is the number of features in S, $|F|$ is the number of features in F, and $[f_{SUi}(x_{SUi})$ - $f_S(x_S)]$

is the difference in the model's output when feature i is added to the coalition S compared to when feature i is not included in the coalition.

To calculate Φ_i, we need to consider all possible coalitions of features that can be formed from the set of all features except i, and compute the marginal contribution of feature i to each coalition. The term $|S|!(|F|-|S|-1)!$ represents the number of ways to order the features in the coalition S, and normalizes the contribution of each coalition.

When dealing with complex models with a large number of features, computing the exact Shapley values can be a daunting task due to the need to consider all possible combinations of features and their contributions, which can result in a computationally expensive process. To address this challenge, the authors of the paper "A Unified Approach to Interpreting Model Predictions" [13] introduce two model-agnostic approximation methods for computing the Shapley values, namely Shapley sampling values [16] and Kernel SHAP.

2.3 A practical study

Breast cancer is a disease that affects many people worldwide, and accurate diagnosis and treatment decisions are crucial for improving patient outcomes. Machine learning techniques, such as LIME and SHAP, have shown promise in predicting the overall survival of breast cancer patients.

The paper "Machine Learning Explainability in Breast Cancer Survival" [10] focuses on this. It aims to create a predictive model of 10-year overall survival (OS) after curative breast cancer surgery using data from the Netherlands Cancer Registry (NCR). LIME and SHAP were applied to explain the obtained model's predictions. The study used NCR data, including demographic, clinical, and pathological data of 46,284 patients with non-metastatic breast cancer who underwent surgery. The final dataset consisted of 31 features. The six best ranked features were age, ratly, rly, ptmm, pts, and grd. Random Forest, XGB, KNN, Artificial Neural Networks, Naive Bayes, and Logistic Regression were experimented with. LIME and SHAP were used to understand the model's predictions. LIME approximates individual predictions of a black box model with a local surrogate model, while SHAP allows global explanation by expressing it as linear functions of features. Consistency of LIME was assessed by applying it to each of the test set's predictions 100 times. The agreement between LIME and SHAP values was tested by comparing their instances' explanations in the test set.

The study's findings were presented through Figure 2, which displayed five randomly selected patients and the weights assigned by LIME to their features. The x-axis represented a particular feature's value, while the y-axis showed the weight assigned by LIME. Positive weights suggested that the feature contributed to survival, while negative weights suggested that it contributed to death. The boxes on the plot displayed each patient's feature values, with their position being the same across all plots. Inconsistencies were indicated by boxes crossing the dotted line at 0 in the figure below.

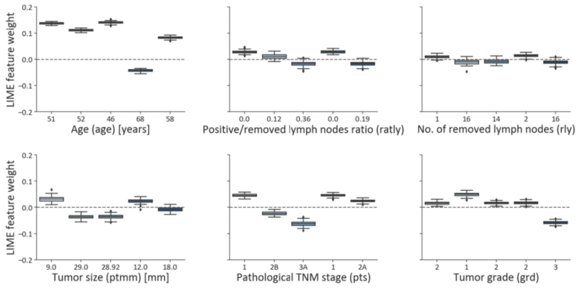
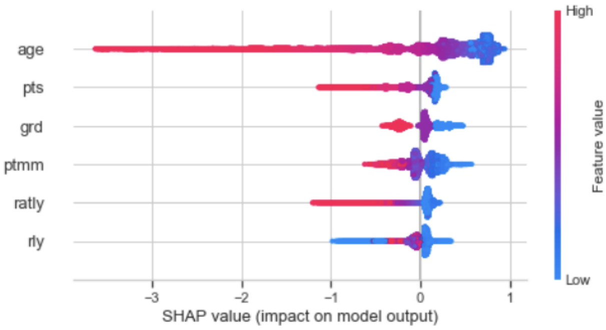
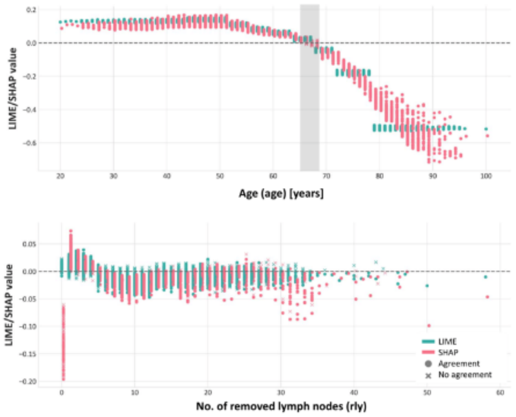


Figure 2: LIME consistency of five representative data instances (i.e. patients) picked at random

The SHAP values of the six features used by the global model are displayed in the figure below. The y-axis represents the features, and the x-axis shows the corresponding SHAP values. The color scale reflects the feature values and ranges from low (blue) to high (red). As with LIME, a positive SHAP value indicates that the feature contributed to survival, while a negative value indicates it contributed to death.



Here below, it is presented an image that combines LIME and SHAP. Although LIME and SHAP values show a similar trends in some features (in the case showed in the image: rly and age), we can also distinguish specific regions of mismatch. Inconsistencies in age, for instance occur around the ages of 65 and 68 (darkened region). This might explain why the model believes age to be a factor in survival or death.



3 Explainability techniques in breast cancer prediction and diagnosis

In this section, we explore various techniques for enhancing explainability in medical image analysis, specifically for breast cancer prediction and diagnosis. The three primary approaches investigated include:

Case-based Ensemble Learning System: A method that combines multiple case-based reasoning (CBR) models to predict breast cancer recurrence, offering explanations through the most similar cases and their features.

Visual Case-Based Reasoning Approach: An approach that employs a visual interface, allowing users to explore similarities between cases and facilitating a better understanding of how each feature contributes to the classification decision.

Explain Individual Classification Decisions: A technique that approximates the classifier with another classifier resembling the Bayes classifier, calculating explanation vectors that provide a quantitative measure of feature importance in complex models and high-dimensional data.

We compare these techniques in terms of their quantitative and qualitative explainability, as well as user study assessment when available. The case-based ensemble learning system and visual case-based reasoning approach primarily focus on qualitative explanations, with the latter also demonstrating strong quantitative explainability. The technique for explaining individual classification decisions emphasizes quantitative explainability but may have limitations in providing qualitative insights. User study assessment is available only for the visual case-based reasoning approach, which received positive feedback on its usability and effectiveness.

By examining these approaches, this report aims to provide valuable insights into the current state of explainable artificial intelligence techniques in the domain of breast cancer prediction and diagnosis, highlighting their advantages and limitations. This investigation can guide future research and development efforts in creating more transparent, understandable, and effective AI-driven tools for healthcare professionals and patients alike.

3.1 Case-based Ensemble Learning System for Explainable Breast Cancer Recurrence Prediction

The authors propose a method that combines case-based reasoning (CBR) and ensemble learning to predict the risk of breast cancer recurrence and explain the reason for the prediction. They use extreme gradient boosting (XGBoost), a state-of-the-art ensemble learning method, to achieve better generalization ability and accuracy. The CBR approach helps doctors understand the underlying decision-making process by providing a case-based interpretation of its prediction. The article discusses the advantages of this approach for improving the quality of doctors' decision making.

3.1.1 Methodology

The methodology involved the following steps:

1. Data preprocessing: Data from 1,286 breast cancer patients who underwent surgery in a Chinese hospital were gathered by the authors. The authors then pre-processed the data by removing missing values and outliers, and normalized it to ensure that all features were of equal importance.

2. Ensemble learning: The authors employed the state-of-the-art XGBoost method to predict the risk of breast cancer recurrence. By training and combining multiple classifiers, XGBoost achieved superior generalization ability and accuracy.

3. Case-based reasoning: The authors utilized

case-based reasoning (CBR) to justify the prediction's reasoning. CBR is a problem-solving approach that leverages prior experiences to tackle new issues. In this case, CBR provides an interpretation of the prediction based on past cases, making it easier for physicians to comprehend the decision-making process.

4. Evaluation: The authors evaluated their system via 10-fold cross-validation and compared it to other advanced methods for predicting breast cancer recurrence. They also made the system available to 32 oncologists and conducted a survey to determine its usefulness as perceived by end users.

5. Discussion, conclusion, and perspectives: The authors discussed the advantages of their approach for improving the quality of doctors' decision making in medical decision support systems. They also highlighted some potential applications of their explainable AI system beyond breast cancer recurrence prediction.

3.1.2 Experimentation and Results

The evaluation results demonstrate that the proposed system displayed exceptional predictive performance. The authors compared their system with other advanced techniques like logistic regression, support vector machine (SVM), random forest, and deep learning for predicting breast cancer recurrence. The study's findings indicated that their system outperformed all other techniques in terms of accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC-ROC).

The authors also provided the system to 32 oncologists and conducted a survey among them to assess its utility as perceived by end users. The results of the survey showed that the oncologists found the system useful and easy to use. They also found the case-based interpretation provided by CBR helpful in understanding the underlying decision-making process.

3.1.3 Advantages and Limitations

Advantages

1. High accuracy: The system displayed outstanding predictive performance, outperforming other advanced methods for predicting breast cancer recurrence.

2. Interpretable: The case-based reasoning approach offers a case-specific explanation for the prediction, facilitating doctors' comprehension of the decision-making process.

3. User-friendly: According to the authors' survey, the system was beneficial and straightforward for oncologists to use.

4. Potential to enhance clinical decision-making: The system may aid physicians in making better decisions in clinical practice, potentially improving patient outcomes.

Limitations

1. Limited data: The authors' data collection was from only one hospital in China, which may limit the generalizability of their findings to other populations.

2. Limited features: The authors utilized a restricted number of features to predict breast cancer recurrence, which may not capture all the relevant information.

3. Limited evaluation: While the authors evaluated their system using 10-fold cross-validation and compared it with other state-of-the-art methods, further evaluation on larger and more diverse datasets is needed to fully assess its performance.

4. Limited scalability: While the proposed system is accurate and interpretable, it may not be scalable to larger datasets or more complex medical problems without significant modifications or extensions.

3.1.4 Conclusion

The proposed system combines ensemble learning with case-based reasoning to provide an accurate and interpretable system for breast cancer recurrence prediction that can help doctors make better decisions in clinical practice. The system displays exceptional predictive performance and outperforms other advanced methods for predicting breast cancer recurrence. The case-based reasoning approach allows for easier understanding of the underlying decision-making process, making the system useful and user-friendly for oncologists. However, there are limitations related to data, features, evaluation, and scalability that should be addressed in future research to further enhance the system's performance and applicability.

3.2 Explainable Artificial Intelligence for Breast Cancer: A Visual Case-Based Reasoning Approach

This paper proposes a breast cancer diagnosis methodology using Case-Based Reasoning (CBR) and visual explanation. The proposed approach combines quantitative and qualitative methods to provide a powerful tool for breast cancer diagnosis that is both accurate and explainable. The researchers used public datasets related to breast cancer to build a knowledge base of similar cases and evaluated the performance of the visual CBR approach against a conventional CBR approach. The results demonstrate that the visual CBR approach is a reliable tool for breast cancer diagnosis, producing both accurate and explainable results.

3.2.1 Methodology

1. Data gathering: Obtain data on breast cancer instances, such as patient demographics, medical history, imaging results, and treatment outcomes, to establish a knowledge base of comparable cases that can aid in future diagnoses.

2. Feature identification: Identify pertinent characteristics from the collected data that can be used to compare and categorize new cases.

3. Similarity calculation: Calculate the similarity between the new case and the cases in the knowledge base using a distance metric such as Euclidean distance or cosine similarity.

4. Case retrieval: Retrieve a set of similar cases from the knowledge base that are most relevant to the new case based on the calculated similarity scores.

5. Visual interface: Display the retrieved cases in a visual interface that allows users to explore both quantitative and qualitative similarities between the query case and each retrieved case.

6. Visual reasoning: Perform visual reasoning to classify the query case based on its similarities with retrieved cases, allowing for fully explainable decision making.

7. Automatic algorithm: Provide an automatic algorithm for classifying new cases based on their similarity scores with retrieved cases.

8. Explanation generation: Generate explanations for why a particular classification decision was made by highlighting relevant features and their contributions to the decision.

3.2.2 Datasets

- The Breast Cancer Wisconsin (BCW) dataset.

- The Mammographic Mass (MM) dataset.

- The Breast Cancer (BC) dataset.

3.2.3 Experiments and Results

The visual CBR approach outperformed the conventional CBR approach in terms of accuracy, precision, and F1 score. Specifically, the visual CBR approach achieved an 85% accuracy rate, compared to the conventional CBR approach's 75%. The precision and recall measures were also higher for the visual CBR approach. User feedback indicated that the interface was user-friendly and helped users understand how each feature influenced the classification decision.

3.2.4 Advantages and Disadvantages

Advantages

1. Accuracy: Higher accuracy than traditional CBR approaches.

2. Explainability: Clear explanations for classification decisions.

3. Usability: User-friendly interface.

4. Adaptability: Solves two different optimization problems using the same metaheuristic.

Disadvantages

1. Limited applicability: Specific to breast cancer diagnosis.

2. Data availability: Relies on the availability of high-quality data.

3. Technical proficiency: Requires expertise in both CBR and data visualization.

3.2.5 Conclusion

Despite its drawbacks, the visual CBR approach's benefits make it a promising tool for enhancing the accuracy and interpretability of breast cancer diagnosis. By combining quantitative and qualitative methods, this approach provides a powerful and explainable decision-making tool for clinical settings. The user-friendly interface and clear visual explanations further contribute to the utility of this approach. Future work may focus on expanding the applicability of this method to other medical conditions and addressing the limitations related to data availability and technical proficiency.

3.3 Explain individual classification decisions

This paper presents a method for explaining individual classification decisions by approximating the classifier with another classifier, $\hat{g}$, which resembles the Bayes classifier. This approximation allows for the calculation of explanation vectors that help understand the classifier's decision on test data points. The results demonstrate the method's effectiveness in terms of accuracy and computational efficiency compared to other state-of-the-art methods, such as LIME and SHAP.

3.3.1 Methodology

The methodology for explaining individual classification decisions involves the following steps:

1. Define local explanation vectors as class probability gradients.
2. Apply the method to Gaussian Process Classification (GPC).
3. Approximate the classifier with another classifier, $\hat{g}$, which resembles the Bayes classifier.
4. Choose an appropriate classifier for $\hat{g}$, such as GPC, logistic regression, or Parzen windows.
5. Estimate local explanations for methods that output a prediction without a direct probability interpretation.

3.3.2 Results and Experimentation

The authors applied their approach to a Support Vector Machine (SVM) classifier for distinguishing digit "2" from digit "8" in the USPS dataset. The method outperforms other methods for explaining individual classification decisions, such as LIME and SHAP, in terms of accuracy and computational efficiency.

3.3.3 Advantages and Disadvantages

The advantages and disadvantages of the proposed method for explaining individual classification decisions are as follows:

Advantages

- Provides a quantitative measure of feature importance through local explanation vectors.
- Can handle complex models and high-dimensional data.
- Flexible and applicable to different types of classifiers.

Disadvantages

- Relies on a good approximation of the classifier, which may not always be possible or accurate.
- Does not capture global properties of the data, only local properties around individual test points.
- Computation time required to estimate local explanations may be high for large datasets.

3.3.4 Conclusion

The proposed method for explaining individual classification decisions offers a promising approach for understanding and interpreting the decisions of complex classifiers. The method's flexibility, quantitative explanation vectors, and ability to handle high-dimensional data make it a valuable tool for practitioners seeking to analyze classification decisions. Future research could focus on improving the approximation of classifiers and incorporating domain knowledge into the explanation process to further enhance the method's utility.

3.4 Discussion

In this section, we investigated three techniques for enhancing explainability in breast cancer prediction and diagnosis using medical image analysis. The techniques explored include case-based ensemble learning systems, visual case-based reasoning approaches, and methods for explaining individual classification decisions. We compared these techniques in terms of their quantitative and qualitative explainability, as well as user study assessment when available.

All three techniques demonstrated their respective strengths and weaknesses. The case-based ensemble learning system and visual case-based reasoning approach primarily focus on qualitative explanations. In contrast, the technique for explaining individual classification decisions emphasizes quantitative explainability but may have limitations in providing qualitative insights. User study assessment was available only for the visual case-based reasoning approach, which received positive feedback on its usability and effectiveness.

This investigation into the current state of explainable artificial intelligence techniques in the domain of breast cancer prediction and diagnosis provides valuable insights into their advantages and limitations. The findings can guide future research and development efforts in creating more transparent, understandable, and effective AI-driven tools for healthcare professionals and patients alike. Furthermore, these techniques can potentially improve clinical decision-making and lead to better patient outcomes.

4 Backpropagation-based techniques

Backpropagation-based techniques refer to a group of methods that use the backpropagation algorithm to explain deep learning models. Widely employed for training deep neural networks, this algorithm aims to assess the sensitivity of output predictions to variations in input features. It accomplishes this by calculating gradients of the output in relation to the inputs. These gradients can be used to indicate the importance or relevance of each input feature to the output prediction. Two of the most known techniques are Grad-CAM and DeepLIFT.

In the following sections, we will begin by delving into the method that underlies Grad-CAM, namely CAM. We will examine CAM and its relevance to understanding Grad-CAM. Additionally, we will explore Score-CAM, a novel method that overcomes certain limitations faced by its predecessors. Lastly we will explore the mechanism behind DeepLIFT.

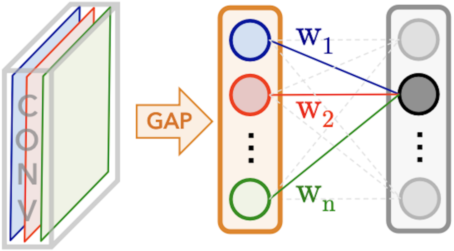
By exploring CAM and Score-CAM together with Grad-CAM and DeepLIFT, we seek to provide a comprehensive overview of these techniques and their applications in the medical field.

4.1 CAM Class Activation Mapping

Class Activation Mapping (CAM)[21] is a visualization technique that identifies the relevant regions of an input image by generating a heatmap that highlights the areas that contributed most. Although CAM is highly effective, it has some significant limitations. Therefore, researchers have developed CAM-based approaches to overcome these limitations and enhance the visual results.

4.1.1 CAM

The idea behind this method is to take advantage of the global pooling layer (GP) present in most CNNs. Before the output layer, the model performs the GP on the convolutional feature maps and use those as features for a fully-connected layer that produces the desired output. Once we have this structure, we can identify the importance of the image regions by projecting back the weights of the output layer on to the convolutional feature maps.



One major limitation of this approach is that it necessitates the inclusion of a global pooling layer in the network. If it is absent, the technique cannot be applied, which considerably restricts its applicability to only those models that contain it. Additionally, CAM can not be applied to networks which use multiple fully-connected layers before the output layer, so fully-connected layers are replaced with convolutional ones and the network is re-trained.

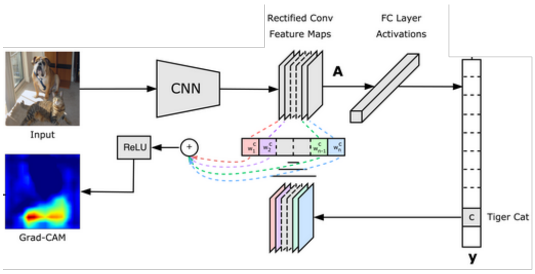
4.1.2 Grad-CAM/++

Gradient-weighted Class Activation Mapping (Grad-CAM)[17] is a class-discriminative localization technique that can generate visual explanations from any CNN-based network *without any architectural requirement* or re-training.

This approach uses the gradient flowing into the final convolutional layer to produce a localization map highlighting the important regions in the image for predicting the specific target.

Unlike CAM, Grad-CAM does not require the presence of a global pooling layers and can be applied to:

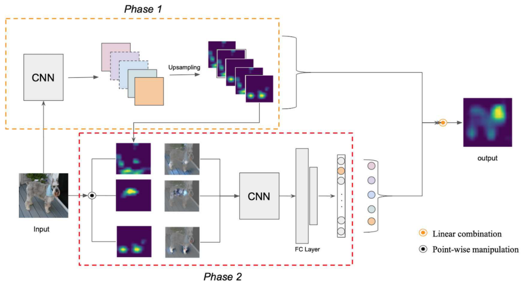
- CNNs with fully-connected layers (e.g. ResNet)
- CNNs used for structured outputs (e.g. captioning)
- CNNs used in tasks with multi-modal inputs (e.g. VQA) or reinforcement learning



The Grad-CAM technique is a powerful and widely used tool but, as pointed out by [20], it has some issues due to the gradient. The gradient of a deep neural network can be **noisy** and also tends to vanish due to **saturation** in sigmoid or the **flat zero-gradient region** in ReLU and this can influence the result of the explanation. Also, since Grad-CAM is linear combination of activation maps, activations with higher weights will have a stronger influence on the result. However, it has been observed that in some cases, the opposite can occur, causing the so called **False Confidence** problem.

4.1.3 Score-CAM

Score-weighted Class Activation Mapping (Score-CAM)[20] is a technique for visualizing the important regions of an image that contribute to a deep neural network's prediction. It has been implemented to overcome the limitations of Grad-CAM. The main difference between Score-CAM and Grad-CAM is in the way they calculate the importance scores for each pixel in the input image. While Grad-CAM uses the gradients of the output class score with respect to the feature maps of the last convolutional layer, Score-CAM directly uses the class activation scores obtained from the last convolutional layer.



Score-CAM is a promising approach and it offers several advantages w.r.t Grad-CAM but it also have some limitations. One major limitation is that it can be computationally expensive, as it requires multiple forward and backward passes through the model, which can increase the inference time. Additionally, it is sensitive to model architecture and hyperparameters, so the performance of Score-CAM can vary depending on the specific architecture used.
Overall, while Score-CAM offers several advantages over Grad-CAM, it is important to consider its limitations and potential sources of error when applying it in practice.

4.1.4 Implementation

Notation We define a CNN as a function $Y = f(X)$ that takes in input $X \in \mathbb{R}^d$ and return a probability distribution Y. We denote A_l as the activation of a given layer l and Y^c as the probability distribution of a given class c. Also, if l is a convolutional layer, we consider A_l^k as the activation map for the $k-th$ channel [20].
Since all the techniques above are based on the same method, we can express a common mathematical formulation:

$$L_t^c = ReLU\left(\sum_k \alpha_k^c A_r^k\right) \tag{1}$$

where $t \in \{CAM, Grad{-}CAM, Score{-}CAM\}$ and r is the reference we take in the activation map (e.g. the entire channel l or only one pixel ij).
It's essential to note that the previous formulation, which includes a ReLU activation, is a specific generalization introduced by the authors of Score-CAM [20], while *the original CAM formulation does not incorporate a ReLU*. Given this formula, the difference between these techniques is the way the scores α_k^c are evaluated:

- **CAM** → $\alpha_k^c = w_{l,l+1}^c[k]$
- **Grad-CAM** → $\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{GAP}} \frac{\partial Y^c}{\partial A_{ij}^k}$
- **Score-CAM** → $a_k^c = C(A_l^k)$

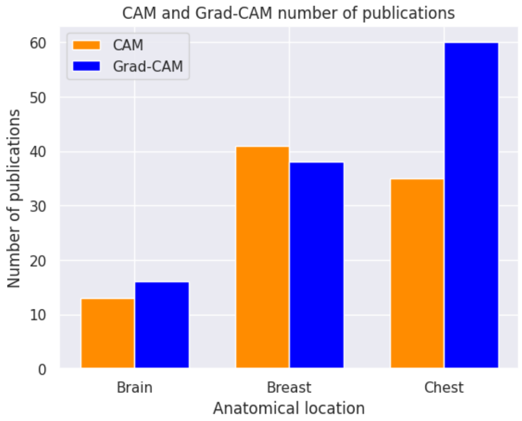
where $C(\cdot)$ is the Channel-wise Increase of Confidence (CIC) defined in [20].

4.1.5 Healthcare applications

Now that we have briefly discussed how certain techniques work, let us shift our focus to their application in the medical field. In recent years, there has been a growing focus on increasing the transparency of medical models and ensuring that they are reliable and accurate. As a result, Explainable Artificial Intelligence (XAI) techniques have become increasingly popular in the medical field. These techniques allow doctors and medical professionals to understand how a particular diagnosis or treatment recommendation was reached, thereby increasing trust and improving patient outcomes.

As we can see in [19] CAM based approaches are used for several "Anatomical locations" and some area where these techniques have been particularly beneficial is in breast, brain and chest cancer detection. Medical imaging techniques like mammography and chest X-rays have traditionally been the primary means of detecting these type of cancer.

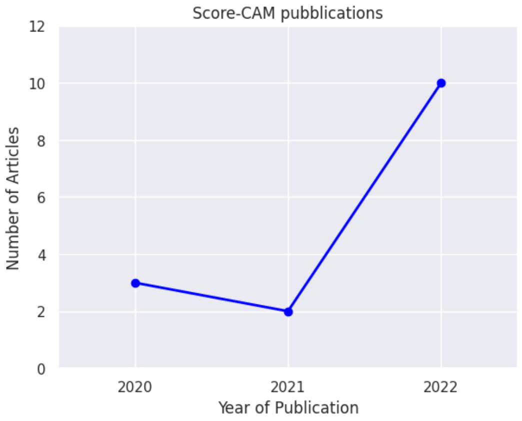
After conducting an advanced search on PubMed, a scholarly literature database in the medical domain, we discovered interesting results that highlight the widespread use and extensive research on the techniques we investigated. These findings underscore the techniques relevance and potential for further advancement in various medical applications. In the table below we show the number of papers for comparison:



To ensure an up-to-date comparison, we conducted a search of all articles published on PubMed between 2018 and 2023.

Since Score-CAM came out in 2020, the articles were not as much as the other methods so we didn't included it in the comparison plot. Despite its current limited use, we believe that this technique holds promise and anticipate that it will gain wider adoption in the coming years especially if we consider his performances (section 4.1.6).

The plot below illustrates a notable increase in the usage of this method over the last 3 years:



4.1.6 Performance

Performance is a critical factor in determining the effectiveness of these techniques, for this type of task we can't use the conventional metrics that we use for classification tasks such as accuracy or F1 score, we need something more specific.

A common metric is the **localization** which refers to the ability of a deep learning model to identify and localize the regions of an input that are most relevant to its output.

Each technique use a different type of evaluation but in order to compare them we can use two very useful metrics introduced by Chattopadhay *et al* [5].

Average drop % This metric relies on the fact that a good explanation map for a class should highlight the most important regions. In order to evaluate this, we first perform the classification task on the original image and then on the explanation map. At this point we compare the confidence of the two classifications on the the same class. The lower is the value the better is the explanation map because a low value indicates that the map highlights a relevant region in the image.

Increase in confidence % This metric measures the number of times in the entire dataset in which the model's confidence increase upon occluding unimportant regions. The higher is the value the better is the explanation map because an high value indicates that the map is giving a low score to less relevant region of the image.

In the following table we can see a comparison made by Wang *et al* [20] between Score-CAM and Grad-CAM according to the metrics defined above:

Metric	**Grad-CAM**	**Score-CAM**
Average Drop(%)	47.8	**31.5**
Average Increase(%)	19.6	**30.6**

4.2 DeepLIFT

DeepLIFT (Deep Learning Important FeaTures), is a method for decomposing the output prediction of a neural network on a specific input by backpropagating the contributions of all neurons in the network to every feature of the input [18].
This technique explains the difference of output from some 'reference' output in terms of the difference of the input from some 'reference' input. *The 'reference' input represents some default or 'neutral' input* that is chosen according to what is appropriate for the problem at hand.

Notation Let t represent some target output neuron of interest and let $x_1, x_2, ..., x_n$ represent some neurons in the intermediate layer or set of layers that are necessary and sufficient to compute t. We consider t^0 the reference activation of t. The difference from the reference activation and the target neuron (difference-from-reference) is defined as $\Delta t = t - t^0$.

Implementation DeepLIFT assign the contribution scores $C_{\Delta x_i \Delta t}$ to Δx_i so that:

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t \tag{2}$$

We refer to this equation as **summation-to-delta** property.
$C_{\Delta x_i \Delta t}$ can be thought of as the amount of difference-from-reference in t that is attributed to the difference-from-reference of x_i.
We can notice two important things:

- **the score is non-zero even when the gradient $\frac{\partial t}{\partial x_i}$ is zero**; this property allow DeepLIFT to address a fundamental limitation of gradients because **a neuron can be meaningful even in the regime where its gradient is zero**
- **the difference-from-reference is continuous**, differently from the gradients which have discontinuous nature that causes sudden jumps in the confidence score, allowing DeepLIFT to **avoid discontinuities** caused by bias terms

Multipliers Given and input neuron x and a target neuron t with difference-from-reference Δx and Δt, we can define the multipliers as:

$$m_{\Delta x \Delta t} = \frac{C_{\Delta x \Delta t}}{\Delta x}$$

If we have $x_1, ..., x_n$ neurons in the input layer and $y_1, ..., y_n$ neurons in an hidden layer, it's possible to formulate the definition of $m_{\Delta x_i \Delta t}$ so that it is consistent with the summation-to-delta property in Eq. 2:

$$m_{\Delta x_i \Delta t} = \sum_j m_{\Delta x_i \Delta y_i} m_{\Delta y_i \Delta t}$$

We refer to this equation as **chain rule for multipliers** property.
Given the multipliers for each neuron to its immediate successors, we can compute the multipliers for any neuron to a given target neuron efficiently via backpropagation [18].

4.2.1 Assigning Contribution Scores

We assume that the reference of a neuron is its activation on the reference input, so the neuron y with input $x_1, x_2, ...$ is given by $y = f(x_1, x_2, ...)$. Consequently, given the reference activations of the input $x_1^0, x_2^0, ...$ of the inputs, we can calculate the reference activation of the output y^0 of the output as:

$$y^0 = f(x_1^0, x_2^0, ...)$$

The choice of a reference input is extremely important; it rely on domain-specific knowledge, and in some cases it may be best to compute DeepLIFT scores against multiple different references.

Positive and Negative Contributions The distinction between positive and negative contributions is essential because we have treat them differently. For every neuron y we will represent the positive and the negative components of Δy respectively as Δy^+ and Δy^- such that:

$$\Delta y = \Delta y^+ + \Delta y^-$$
$$C_{\Delta y \Delta t} = C_{\Delta y^+ \Delta t} + C_{\Delta y^- \Delta t}$$

In addition to the chain rule defined in section 4.2, the authors introduce three rules for computing the contribution of input features based on their activation differences with respect to a reference input:

- **Linear Rule:** This rule *applies to Dense and Convolutional layers* (excluding nonlinearities). Given a linear function y with input x_i such that $y = b + \sum_i w_i x_i$ we can define

$\Delta y = b + \sum_i w_i \Delta x_i$. Now we can define also the positive and the negative parts of Δy as:

$$\begin{aligned}\Delta y^+ &= \sum_i 1\{w_i \Delta x_i > 0\} w_i \Delta x_i \\ &= \sum_i 1\{w_i \Delta x_i > 0\} w_i \left(\Delta x_i^+ + \Delta x_i^-\right)\end{aligned}$$

and

$$\begin{aligned}\Delta y^- &= \sum_i 1\{w_i \Delta x_i < 0\} w_i \Delta x_i \\ &= \sum_i 1\{w_i \Delta x_i < 0\} w_i \left(\Delta x_i^+ + \Delta x_i^-\right)\end{aligned}$$

- **Rescale Rule:** *This rule applies to nonlinear transformations that take a single input* (e.g. ReLU, sigmoid).
  Let neuron y be a nonlinear transformation of the input x such that $y = f(x)$. Since y has only one input, we have by summation-to-delta that $C_{\Delta x \Delta y} = \Delta y$ and $m_{\Delta x \Delta y} = \frac{\Delta y}{\Delta x}$.
  For the Rescale Rule we can set the positive and the negative components Δy^+ and Δy^- as follow:

  $$\begin{aligned}\Delta y^+ &= \frac{\Delta y}{\Delta x}\Delta x^+ = C_{\Delta x^+ \Delta y^+} \\ \Delta y^- &= \frac{\Delta y}{\Delta x}\Delta x^- = C_{\Delta x^- \Delta y^-}\end{aligned}$$

  and base on this, consequently, we have:

  $$m_{\Delta x^+ \Delta y^+} = m_{\Delta x^- \Delta y^-} = m_{\Delta x \Delta y} = \frac{\Delta y}{\Delta x}$$

  We use the gradient instead of the multiplier when x is close to its reference x^0 to avoid numerical instability issues caused by having a small denominator.

- **RevealCancel Rule:** Rescale rule improves using gradients but there are still some situations where it can provide misleading results. If we have two identical inputs, using the Rescale rule, all importance would be assigned to only one of them; this can obscure the fact that both inputs are relevant for the min operation.
  The idea is to treating the positive and negative contributions separately.
  ”In other words, we set Δy^+ to the average impact of Δx^+ after no terms have been added and after Δx^- has been added, and we set Δy^- to the average impact of Δx^- after no terms have been added and after Δx^+ has been added. This can be thought of as the Shapely values of Δx^+ and Δx^- contributing to y.”

4.2.2 Healthcare applications

Unlike CAM-based techniques that are limited to visualizing the features learned by CNNs, DeepLIFT is a more versatile technique that can be applied to all types of deep learning models, including CNNs, recurrent neural networks (RNNs), and feedforward neural networks (FNNs). As such, DeepLIFT can be used for a variety of applications, including natural language processing (NLP) and computer vision, allowing it to provide a comprehensive view of how the deep learning model is making decisions. This flexibility of DeepLIFT makes it a valuable tool for understanding the inner workings of deep learning models, and it has been used to gain insights into a variety of clinical domains.
There are various applications in the healthcare field, which can be broadly divided into two categories:

- **Image Analysis** → interpret deep learning models that analyze medical images such as X-rays and MRIs [12]. By attributing importance scores to each pixel or region of the image, DeepLIFT can help clinicians understand which parts of the image are most relevant for making the diagnosis. This information can aid in decision-making, improve the accuracy of diagnoses, and reduce the risk of misdiagnosis

- **Data analysis** → interpret deep learning models trained on electronic health records (EHRs) or other types of medical data. EHRs contain a wealth of information about patients, including their medical history, laboratory test results, and clinical notes. Deep learning models trained on this data can predict outcomes such as the risk of developing a particular disease or the likelihood of hospital readmission. DeepLIFT can help clinicians understand which factors or features in the EHR are most important in making these predictions, which can aid in treatment planning and improve patient outcomes.

However, it can be challenging to interpret the output of these models and understand which features or regions of the image were crucial in making the diagnosis and it can be challenging to understand which variables or features in the EHR were most important in making these predictions.

4.3 Experimental Analysis

This section aims to evaluate the localization ability of Grad-CAM on a CNN that has been fine-tuned specifically for brain tumor classification. We will begin by presenting the experimental setup in detail, followed by an analysis of the results obtained through this technique.

Setup We conducted an experiment with the objective of testing the impact of utilizing Grad-CAM in the context of our study. To achieve this, we applied Grad-CAM to ResNet50 [9], a widely used convolutional neural network architecture.

However, due to ResNet50 being pre-trained on ImageNet, a dataset that does not include any medical images, we needed to fine-tune it to adapt it to our specific task. To accomplish this, we utilized a dataset known as the "Brain MRI images for brain tumor detection" [4] which is specifically designed for brain tumor image classification tasks. The dataset comprises a diverse collection of brain magnetic resonance imaging (MRI) scans and is annotated with two distinct classes: "Normal" and "Tumor." These labels allowed us to train ResNet50, enabling it to learn the necessary features for accurate brain tumor classification.

After successfully fine-tuning ResNet we proceeded to apply Grad-CAM -using an implementation provided by [6]- to highlight the regions it considered important for making tumor classification decisions.

Results In this section, we will discuss the performance of our models and the improvements achieved throughout the experimentation process.

We started our investigation by employing a basic fine-tuned model, which exhibited an accuracy of 80%. Despite the initial moderate success, we recognized the need for further enhancements to achieve more accurate and reliable tumor localization. As a result, we continued to refine our approach and managed to significantly improve the performance, ultimately attaining an accuracy of 92%.

A crucial finding from our comparative analysis was the noticeable difference in the localization abilities of the two models. The first model, while reasonably accurate, tended to primarily focus on the central region of the images. This limitation posed challenges in accurately identifying tumors in different positions within the images. In contrast, the second model exhibited significant advancements in localizing tumors throughout the entire image. It effectively highlighted tumors in various positions, indicating a superior capability for localization compared to the initial model.

Based on the outcomes of our study, we can infer that achieving high performance models is crucial for accurate tumor localization. A model with superior localization capabilities enables medical practitioners to precisely identify and analyze tumors, facilitating more effective diagnosis and treatment planning.

In summary, the obtained results of our study were found to be satisfactory, indicating the effectiveness of the Grad-CAM technique in localizing tumors in medical images. The stark contrast between the localization abilities of the initial and refined models emphasizes the importance of model performance for achieving accurate tumor localization. These findings contribute to the broader understanding of the role of deep learning techniques and the potential of Grad-CAM in the field of medical imaging analysis.

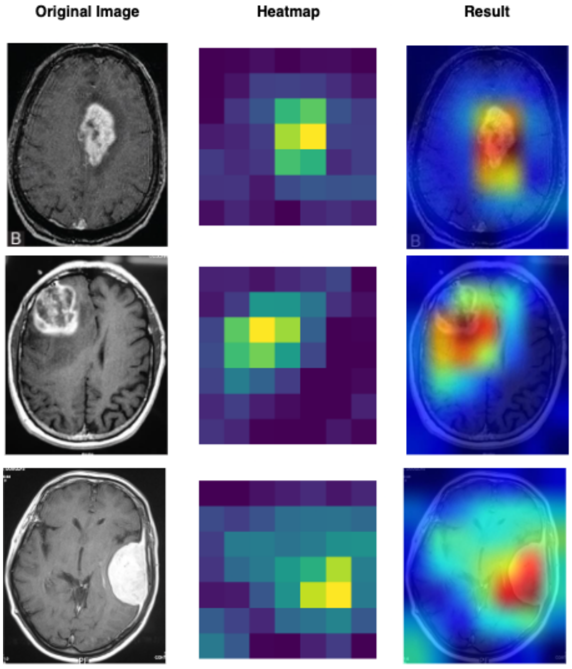


Figure 3: Visualization of results achieved with the Grad-CAM method applied to the higher performance model. The code is available on *https://github.com/daniele-marini/Grad-CAM*

5 Conclusion

In conclusion, this research paper has explored various techniques in explainable artificial intelligence (XAI) applied to the medical field, with a focus on cancer prediction and diagnosis. Specifically, the techniques discussed include SHAP, LIME, case-based ensemble learning systems, visual case-based reasoning approaches, and CAM, Grad-CAM, Score-CAM, and DeepLIFT. Each technique contributes to enhancing the interpretability and transparency of machine learning models in different ways.

The utilization of SHAP and LIME techniques has showcased their ability to provide valuable insights into the decision-making process of complex machine learning models. By quantifying feature importance scores and generating local explanations, these techniques enable healthcare professionals to understand the factors influencing individual predictions, thereby improving the trust and acceptance of AI models in critical medical applications. The interpretability offered by SHAP and LIME not only aids in verifying the reliability of the model but also enhances clinical decision-making by offering a deeper understanding of the underlying patterns and relationships within the breast cancer data.

Moreover, our investigation into case-based ensemble learning systems and visual case-based reasoning approaches has shed light on their strengths and weaknesses in terms of explainability. These techniques emphasize qualitative explanations, allowing healthcare professionals to gain insights into the reasoning behind predictions. On the other hand, the technique for explaining individual classification decisions provides a quantitative perspective, but it may have limitations in providing detailed qualitative insights. The positive user study assessment of the visual case-based reasoning approach further underscores its usability and effectiveness, offering a promising direction for future research and development in creating more transparent, understandable, and effective AI-driven tools for breast cancer prediction and diagnosis.

Additionally, the comparative analysis of CAM, Grad-CAM, Score-CAM, and DeepLIFT techniques has underscored their collective significance in providing visual explanations and interpretability for deep learning models. These techniques have demonstrated their value in various medical applications, including disease detection, tumor segmentation, and anomaly localization. The visual explanations generated by CAM-based methods, coupled with the in-depth analysis enabled by DeepLIFT, foster transparency and trust in AI-driven medical systems. This collaboration between clinicians and algorithms enables more accurate diagnoses, personalized treatments, and advancements in precision medicine.

In summary, the integration of a range of XAI techniques in the medical field, encompassing SHAP, LIME, case-based ensemble learning systems, visual case-based reasoning approaches, CAM, Grad-CAM, Score-CAM, and DeepLIFT, holds immense promise in improving the transparency, trust, and interpretability of AI systems. The marriage of artificial intelligence with human expertise fosters a collaborative environment, where the strengths of both clinicians and algorithms are harnessed to enhance breast cancer care. Through continued research and refinement, these techniques can usher in an era of accurate diagnoses, personalized treatments, and significant advancements in precision medicine, ultimately benefiting patients worldwide.

References

[1] 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.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[3] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

[4] Navoneel Chakrabarty. Brain mri images for brain tumor detection. `https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection`, 2018.

[5] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on ap-*

plications of computer vision (WACV), pages 839–847. IEEE, 2018.

[6] François Chollet. Grad-cam class activation visualization. `https://github.com/keras-team/keras-io/blob/master/examples/vision/grad_cam.py`, 2020.

[7] Damien Garreau and Dina Mardaoui. What does lime really see in images?, 2021.

[8] Dongxiao Gu, Kaixiang Su, and Huimin Zhao. A case-based ensemble learning system for explainable breast cancer recurrence prediction. *Artificial Intelligence in Medicine*, 107:101858, 2020.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Tom Jansen, Gijs Geleijnse, Marissa Van Maaren, Mathijs P Hendriks, Annette Ten Teije, and Arturo Moncada-Torres. Machine learning explainability in breast cancer survival. In *Digital Personalized Health and Medicine*, pages 307–311. IOS Press, 2020.

[11] Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud, and Brigitte Séroussi. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial intelligence in medicine*, 94:42–53, 2019.

[12] Alina Lopatina, Stefan Ropele, Renat Sibgatulin, Jürgen R Reichenbach, and Daniel Güllmar. Investigation of deep-learning-driven identification of multiple sclerosis patients based on susceptibility-weighted images using relevance analysis. *Frontiers in neuroscience*, 14:609468, 2020.

[13] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[14] 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.

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning, 2016.

[16] Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. The shapley value in machine learning. *arXiv preprint arXiv:2202.05594*, 2022.

[17] 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.

[18] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[19] Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, page 102470, 2022.

[20] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

[21] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.