Journal of Information Systems Engineering and Management
2025, 10(29s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

An Explainable Deep Learning Model Combining Integrated Gradients, GradientSHAP and Occlusion for Breast Cancer Detection

S. Ranjana[1], A. Meenakshi[2]

[1] Department of Computer Science and Applications, SRM Institute of Science and Technology Vadapalani Campus, Chennai, Tamil Nadu, India. rs2363@srmist.edu.in

[2]Department of Computer Science and Applications, SRM Institute of Science and Technology Vadapalani Campus ,Chennai, Tamil Nadu, India. meenaksa@srmist.edu.in

ARTICLE INFO

Received: 31 Dec 2024
Revised: 20 Feb 2025
Accepted: 28 Feb 2025

ABSTRACT

Breast cancer, a prominent cause of female mortality, underscores the critical need for tim intervention. Researchers and professionals are continually developing models for early breast cancer detection, yet the aspect of interpretation remains underexplored. This research is motivated by the desire to visually interpret breast cancer diagnoses for pathologists and even individuals with a layman's interest. The study utilizes the BreakHis dataset from the Kaggle and employs a ResNet50 model through deep learning to classify breast tumors as either malignant or benign. The model achieved a notable 96.84% accuracy in testing, surpassing results obtained by other researchers using similar explainable deep learning methodologies.

Furthermore, this research goes beyond mere classification by employing eXplainable Artificial Intelligence (XAI) techniques—Integrated Gradient (IG), GradientShap (GS), and Occlusion with CNN Model to interpret breast cancer diagnoses from histopathological images. These techniques provide insights into why a specific histopathological image is categorized as benign or malignant. Among the deep learning technique CNN and with explainable AI techniques XAI, XAI stands out for its superior predictive results based on visualization Unlike earlier studies, this research not only classifies histopathological images but also offers transparent reasons for its outcomes, enhancing the understanding of breast cancer diagnoses.

Keywords: Breast Cancer, Resnet5,XAI,Deep learning,GradientShap,Occlusion,Integrated Gradient.

INTRODUCTION

Cancer, a widespread term encompassing abnormal cell growth in the human body, stands as a leading global cause of death. In various forms of cancer, numerous body tissues undergo uncontrolled division and proliferation. Cancer can originate in nearly any part of the human body. Typically, human tissues undergo regulated growth and division to generate new tissues as needed. Under normal circumstances, aging cells are meant to undergo programmed cell death and be replaced by new ones. However, in the case of cancer, this orderly process is disrupted: older cells, which should naturally expire and be replenished, persist, while new cells, unnecessary for the body, continue to form [1] and abnormal growth of cells is known as tumors and Solid tumors, comprised of masses of cells, manifest in various types of cancer. Tumors can be categorized as either benign or malignant [2] A benign tumor is not cancerous, exhibiting slow growth and an inability to spread to other body parts. Surgical removal often prevents its recurrence. In contrast, malignant tumors are cancerous, featuring uncontrolled cell growth and the ability to spread to other areas of the body. According to assessments by the International Agency for Research on Cancer, 2012 witnessed the diagnosis of 14.1 million new cancer cases, with 8.2 million fatalities globally [3] Several methods utilized for detecting breast cancer encompass Mammography, Magnetic Resonance Imaging (MRI) Scans, Computed Tomography (CT) Scans, Ultrasound, Nuclear Imaging, and others. Among the diverse types of cancers, breast cancer has a higher incidence in women[4]. Several years ago, numerous breast cancer diagnostic models were proposed by researchers employing machine learning techniques. However, these models lacked explainability, especially in the context of the often opaque outcomes generated by "black-box" approaches like deep learning, leading to confusion among pathologists[5]. Physicians, particularly in medical systems, are hesitant to embrace

black-box models as they prefer to comprehend the rationale behind system recommendations. Hence, the need for an explainable model to address this challenge becomes apparent[6].

Previous research has highlighted the difficulty physicians face in providing clear reasons for machine-generated decisions during diagnosis, primarily due to the opaque nature of the models. This lack of transparency has negatively impacted decision-making[7]. To address this gap, there is a call for exposing or explaining the internal operations of each model layer, offering insight into the decision-making process. This visual representation is crucial for the effective interpretation of neural network predictions, especially in the context of diagnosing breast cancer. Such an approach aims to make the system more user-friendly, trusted, and interactive between physicians and machine[8]
The accuracy of medical diagnosis is assessed based on the reliability and consistency of diagnoses derived from clinical variables collected by pathologists. The traditional manual diagnostic tools, allowing pathologists to apply their domain expertise, clash with the introduction of deep learning in clinical decision-making due to its opaque analytics mechanism. This conflict has prompted the automation of clinical diagnostic workflows using explainable deep learning methods. The primary objective of this research is to leverage explainable AI techniques to elucidate the reasons behind breast cancer diagnoses using sampled histopathological images. The study develops an explainable breast cancer prediction system to enhance clinical transparency. To achieve this goal, the research includes the design, implementation, and evaluation of a model based on explainable AI, with a subsequent visualization and interpretation of model decisions using relevant techniques.
Explainable AI (XAI) refers to the development of artificial intelligence models and systems that provide transparent and understandable insights into their decision-making processes. The goal is to make AI systems more interpretable for humans, allowing users to comprehend why a particular decision or prediction was made[9]. Here are some approaches and models within the realm of Explainable AI:

Linear Models:
Linear models, such as linear regression or logistic regression, are inherently interpretable. The relationships between input features and the output are linear and can be easily understood.

Decision Trees:
Decision trees are a type of model that makes decisions based on a sequence of questions or conditions. The decision paths can be visualized, making it easy to understand the decision-making process.

Rule-Based Models:
Rule-based models, including systems like expert systems, formulate decisions as a set of rules. These rules are typically in an "if-then" format, making them interpretable.

LASSO Regression:
LASSO (Least Absolute Shrinkage and Selection Operator) is a regression technique that includes a penalty term, encouraging the model to use fewer features. This can lead to a more interpretable model by emphasizing the most relevant features.

Anchors:
Anchors are a type of rule-based explanation system that focuses on finding simple, understandable conditions that "anchor" the prediction.
Local Interpretable Model-agnostic Explanations (LIME):
LIME is a technique that provides locally faithful explanations for the predictions of any machine learning model. It perturbs the input data and observes the changes in predictions, creating a locally interpretable model.

SHAP (SHapley Additive exPlanations):

SHAP values are based on cooperative game theory and provide a way to fairly allocate a value to each feature, indicating its contribution to a particular prediction.

Counterfactual Explanations:

Counterfactual explanations generate instances of input data that, when fed into the model, would change the prediction. This helps users understand how small changes in input features affect the output.

Interpretable Neural Networks:

Some efforts focus on designing neural network architectures that inherently lend themselves to interpretability, making the decision-making process more transparent.

Symbolic AI:

Symbolic AI techniques involve the use of symbolic reasoning and logic, providing a rule-based and transparent approach to decision-making.

LITERATURE REVIEW

Researchers worldwide have increasingly applied neural network methods to medical image analysis, yielding promising outcomes. Numerous studies have explored the use of Deep Learning techniques for diagnosing Breast Cancer in histopathological images, demonstrating the potential of artificial intelligence technology.

Authors	Methodolgy	Advantage	Disadvantage
I-S. Jung, D. Thapa, and G-N. Wang [10]	Multi-Layer perceptron (MLP), using back propagation	Accuracy 82%	Lacks model explainability.
D. M. Vo, N.-Q. Nguyen and S.-W. Lee [11]	CNN and boosting tree classifier	Highly efficient for feature extraction and learn all level discriminant features of an input medical image,	Poor in Bio imaging datasets when used on single Image Classifier-Black box model
S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider and R. Deters[12]	Multi model ensemble method for a multilayer perceptron classifier	Highly accurate and does automatic feature extraction.	Lack transparency and is not suitable for small sized dataset.
Y. Qiu , S. Yan , R. R. Gundreddy , Y. Wang , S. Cheng , H. Liu and B. Zheng[13]	Deep learning-CNN	feature extraction with high accuracy	dependent on the Region of Interest (ROI)
S. Pratiher and S. Chattoraj[14]	Ensemble of histological hashing and class-specific manifold learning	High Accuracy, speed in detection and suitability for all both large and small sized data	Lacks model transparency

Table 1 Literature review of models for breast cancer classification

The methodology employed in this study involves the utilization of "Deep Learning." Data set used from kaggle repository, histhopathological images comprising a total of 7,909 images of breast tumor tissue. This dataset encompasses 2,480 benign samples and 5,429 malignant samples.The pursuit of visually interpreting histopathological breast cancer prediction results for the purpose of clinical transparency led to the following sequential steps:

1. Obtain a curated BreakHis dataset.
2. Divide the data into three sets:
   - Training dataset (5,005)
   - Test data (791)
   - Validation dataset (2,113)
3. Apply data augmentation
4. Train a ResNet50 pretrained model using the transfer learning concept.
5. Validate the model's accuracy with sample data.
6. Visualize and interpret model decisions using Integrated Gradient, GradientShap, and Occlusion analyses from PyTorch Captum.

To accomplish this objective, the following specific goals were pursued:
- Perform data augmentation.
- Train a ResNet pretrained model using BreakHis images.
- Assess the model's accuracy with sample data.
- Visualize and interpret model decisions using Integrated Gradient, GradientShap, and Occlusion techniques.

3.1 The current systems, as indicated by the existing literature, are breast cancer prediction models based on deep learning, lacking explainability. They can only determine whether a breast tumor is malignant or benign, providing no insights into the reasons behind the predictions. These systems face several shortcomings, including:

(i) Limited understandability of the model predictions, confined to the developers alone.
(ii) Lack of accountability for the model in cases of predictive errors.
(iii) Mistrust of the model by pathologists due to its lack of explainability.
(iv) Non-scalability, meaning that the model's performance cannot be easily improved since its predictive capabilities cannot be assessed.

Algorithms used for classification

Integrated Gradient (IG) is an explanation method used in machine learning and artificial intelligence to interpret the predictions of a model. It is particularly employed in deep learning models, such as neural networks. IG aims to provide insights into the contribution of each feature or input variable to the model's output by integrating the gradients along the path from a baseline or reference input to the actual input.

The process involves calculating the gradients of the model's output with respect to the input features at multiple points along the path between the baseline and the input. These gradients are then integrated to obtain a comprehensive understanding of how each feature influences the model's decision.

IG is valued for its ability to offer interpretable explanations for complex models, especially those considered "black-box" models, like deep neural networks. By attributing importance scores to input features, IG helps users understand the reasons behind a model's predictions, promoting transparency and trust in AI systems. It finds applications in various domains, including medical diagnosis, where understanding the reasoning behind predictions is crucial.

GradientShap is another explanation method used in machine learning and interpretability, particularly for understanding the predictions of complex models, such as deep neural networks. It falls under the umbrella of Shapley value-based explanation techniques. Shapley values, derived from cooperative game theory, represent the average contribution of a feature to all possible combinations in a predictive model.[15]

GradientShap extends the Shapley value concept to deep learning models. It calculates the Shapley values by considering the average of the model's gradients for each feature across all possible permutations of the input features. In other words, it evaluates the impact of each feature by assessing how its presence or absence influences the model's output.

This method is particularly useful for attributing contributions to individual features in a prediction, offering insights into why a model made a specific decision. It contributes to the interpretability of complex models, providing a more

nuanced understanding of feature importance and facilitating trust in AI systems, especially in critical domains like healthcare.

GradientShap is generally applicable to various machine learning tasks, including breast cancer classification using deep learning models. Below is a simplified explanation of how GradientShap could be applied to a breast cancer classification task:

Model Training:

Train a deep learning model, such as a neural network, on a dataset containing histopathological images of breast tumor tissue. The model should be designed for binary classification, distinguishing between benign and malignant tumors.

Preprocessing:

Preprocess the histopathological images and prepare them for input into the trained model.

Baseline Image:

Choose a baseline image as a reference point for the Shapley value calculation. This could be an image with certain characteristics that are considered neutral or typical.

Gradient Calculation:

For each input image, calculate the gradients of the model's output with respect to each pixel in the image. This is done by backpropagating the gradients through the network.

Shapley Value Calculation:

Average the gradients across multiple images, each with different combinations of features present or absent. The Shapley values represent the average contribution of each pixel to the model's prediction.

Attribution Maps:

Create attribution maps or heatmaps that visualize the Shapley values for each pixel in the input images. These maps highlight regions that significantly contribute to the model's decision.

Interpretation:

Analyze the attribution maps to understand which regions of the histopathological images are influential in the model's classification. This can provide insights into the features the model relies on for distinguishing between benign and malignant tumors.

Occlusion

Occlusion is an interpretability technique used in machine learning and computer vision to understand the importance of different regions in an input data point, often an image. It involves systematically covering or "occluding" parts of the input and observing the impact on the model's prediction.

Explanation of how occlusion works:

1. **Input and Model:**
   - Given an input, typically an image in the context of computer vision tasks like image classification, and a trained machine learning model.

2. **Sliding Window:**
   - A small window or patch is systematically moved across the input, covering different portions at each step. The model's prediction is recorded for each occluded or covered region.
3. **Prediction Impact:**
   - By comparing the model's predictions with and without occlusion, it is possible to understand which regions have a significant impact on the model's decision. If covering a certain region leads to a substantial change in the prediction, it indicates that the occluded region is crucial for the model's decision.
4. **Heatmaps or Maps of Importance:**
   The results can be visualized as heatmaps or maps of importance, highlighting the regions that are most influential in the model's decision. Darker regions in the heatmap may represent areas with higher importance.

In the context of breast cancer classification using histopathological images, occlusion analysis can help identify the critical features or structures in the images that contribute to the model's decision regarding tumor malignancy

RESULTS

AIM	MODEL	ACCURACY	MODEL RESULT EXPLAINED?	EXPLAINABLE ALGORITHM(S) USED
To classify & visually interpret breast histopathological images.	CNN	96.84%	Yes	Integrated Gradient, GradientShap & Occlusion
Classification of breast cancer images.[16]	CNN	91.2%	No	Nil

Table 2 Comparing model with explainable algorithms

DISCUSSION

Table 1 has shown the comparison of model used with and without explainable algorithms for breast cancer classification using CNN models. In this study, the BreakHis dataset was employed, utilizing a deep learning method, specifically a Convolutional Neural Network (CNN), to categorize breast tumors as either benign or malignant. Additionally, Explainable Artificial Intelligence (XAI) techniques, including Integrated Gradient (IG), GradientShap (GS), and Occlusion, were applied to provide visual explanations for the presence of cancer labels in histopathological images. The interpretation of these images using XAI techniques was based on analyzing color variation and location. A higher concentration of black or green coloration in an image indicated a higher likelihood of malignancy or benignity. Upon comparing the three interpretable algorithms, it was observed that Occlusion demonstrated superior visualization capabilities compared to Integrated Gradient and GradientShap, exhibiting less noise and making it more accessible for pathologists to reach conclusions. The experiment achieved an impressive validation accuracy of 96.84%, surpassing related literature due to its robust predictive accuracy and the clarity of visual explanations. This research contributes to the existing knowledge by enhancing the transparency of breast cancer diagnosis results through the implementation of explainable AI techniques.

REFRENCES

[1] P. Dalerba , R. W. Cho and M. F. Clarke (2007). Cancer Stem Cells: Models and Concepts. Annual Review of Medicine. Vol. 58:267-284. doi: 10.1146/annurev.med.58.062105.204854. Retrieved May 10, 2022 from https://www.annualreviews.org/doi/10.1146/annurev.med.

[2] Y. Brazier (2012). Medical News Today. Retrieved April 15th, 2022 from http:/www.medicalnewstoday.com/articles/249141

[3] National Cancer Institute (2012).Cancer Statistics. Retrieved June 25, 2022 from http://www.cancer.gov/about-cancer/what-is -cancer/statistics.

[4] R. L. Siegel , K. D. Miller and A. Jemal (2017). Cancer statistics, 2017. CA: A Cancer Journal of Clinicians.. Volume67, Issue1 January/February 2017, Pages 7-30. doi: 10.3322/caac.21387. Retrieved April 25, 2022 from https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21387

[5] S. Saini and R. Vijay (2014). "Optimization of Artificial Neural Network Breast Cancer Detection System based on Image Registration Techniques." International Journal of Computer Applications (IJCA Journal) , 105 (14): 26–49 . doi:10.5120/18447-9837 Retrieved April 20, 2022 from https://research.ijcaonline.org/volume105/number14/pxc3899837.pdf

[6] X. Wang and O. Gotoh (2009). "Microarray-Based Cancer Prediction Using Soft Computing Approach". Cancer Informatics 2009:7 123–139. https://doi.org/10.4137/CIN.S2655. Retrieved April 22, 2022 from https://journals.sagepub.com/doi/pdf/10.4137/CIN.S2655.

[7] A. Moxey , J. Robertson , D. A. Newby, I. Hains, M .Williamson and S. A Pearson (2010) "Computerized clinical decision support for prescribing: Provision does not guarantee uptake". Journal of the American Medical Informatics Association 17(1):25-33. doi: 10.1197/jamia.M3170. Retrieved June 10, 2022 from https://www.researchgate.net/publication/40907393

[8] P. Khosravi , E. Kazemi , M. Imielinski , O. Elemento and I. Hajirasouliha (2018). Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. EBioMedicine 27(2018), 317–328 https://doi.org/10.1016/j.ebiom.2017.12.026. Retrieved April 21, 2022 from https://reader.elsevier.com/reader/sd/pii/S2352396417305078?token

[9] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula and Matija Snuderl, (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature Medicine (24) 1559–1567. Retrieved July 10, 2022 from https://www.nature.com/articles/s41591-018- 0177-5

[10] I-S. Jung, D. Thapa, and G-N. Wang (2005). "Neural Network Based Algorithms for Diagnosis and Classification of Breast Cancer Tumor"

[11] Y. Hao et al. (Eds.): CIS 2005, Part I, LNAI 3801, pp. 107–114, 2005 Computational Intelligence and Security. CIS 2005. Lecture Notes in Computer Science(), vol 3801. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11596448_15 Retrieved April 22, 2022 from https://page-one.springer.com/pdf/preview/10.1007/11596448_15.

[12] D. M. Vo, N.-Q. Nguyen and S.-W. Lee (2019). Classification of breast cancer histology images using incremental boosting convolution networks. ELSEVIER Information Sciences. Volume(482),123-138. https://doi.org/10.1016/j.ins.2018.12.089 Retrieved July 12, 2022 from https://www.sciencedirect.com/science/article/abs/

[13] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider and R. Deters (2019). "Classification of Histopathological Biopsy Images Using Ensemble of Deep Learning Networks" https://doi.org/10.48550/arXiv.1909.11870. Retrieved July 12, 2022 from https://arxiv.org/abs/1909.11870

[14] S. Pratiher and S. Chattoraj (2019). Diving Deep onto Discriminative Ensemble of Histological Hashing & Class-Specific Manifold Learning for Multi-class Breast Carcinoma Taxonomy. Conference: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE ICASSP). Retrieved January 10, 2022 from https://www.researchgate.net/publication/3327906, doi: 10.1109/ICASSP.2019.8683856

[15] Francis A. U. Imouokhome, Osehi Grace Ehimiyein and Fidelis Odinma Chete(2023). Diagnosis and Interpretation of Breast Cancer Using Explainable Artificial Intelligence.NIPES Journal of Science and Technology Research 5(2) 2023 pp. 102-123 ISSN-2682-5821

[16] Ranjana, S., & Meenakshi, A. (2024, August). Breast Cancer Detection Using Autoencoder with Convolutional Neural Network. In 2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI) (pp. 1391-1399). IEEE. DOI:10.1109/ICoICI62503.2024.10696668