Educational Administration: Theory and Practice
2024, 30(5), 4999-5009
ISSN: 2148-2403
https://kuey.net/

Research Article

Breaking The Black Box: Heatmap-Driven Transparency To Breast Cancer Detection With Efficientnet And Grad CAM

Aditi Kajala[a]*, Sandeep Jaiswal[a], Rajesh Kumar[b]

[a]School of Engineering and Technology, Mody University of Science and Technology, Lakshmangarh, 332311, India
[b]Electrical Engineering, Malaviya National Institute of Technology, Jaipur, 302017, India

Citation: Aditi Kajala et al. (2024), Breaking The Black Box: Heatmap-Driven Transparency To Breast Cancer Detection With Efficientnet And Grad CAM, *Educational Administration: Theory And Practice*, 30(5), 4999-5009, Doi: 10.53555/kuey.v30i5.3738

ARTICLE INFO

ABSTRACT

Breast cancer is a significant global health concern, with the manual diagnostic process being time-consuming. The introduction of Computer-Aided Diagnosis (CAD) has emerged as a promising solution, facilitating quicker and more accessible assessments. However, concerns persist regarding the trustworthiness of these automated systems, particularly deep learning models, due to their inherently black-box nature. Transparency and interpretability are crucial elements, necessitating methods to visualize and comprehend the decision-making process of the model. This research aimed to enhance the transparency and interpretability of deep learning models for breast cancer diagnosis. The focus was on developing a method to highlight prominent areas of histopathology slides using heatmaps. The "Histopathology Cancer Detection (HCD)" dataset was used in the investigation. Eight EfficientNet models were examined for fine-tuning, and a feature extractor for binary classification. The optimized model is further utilized to get the output of any particular layer or block of the model with GradCAM (Gradient-weighted Class Activation Mapping). Heatmaps are produced to show the area of the picture that contributed most to the classification. Notably, the model architecture remained unchanged to maintain diagnostic accuracy, while the introduction of heatmaps aimed to provide additional insights into the decision-making process. To validate the effectiveness of the proposed approach, human validation was conducted. Domain experts were presented with histopathology images along with the model-generated heatmaps. The purpose of the questionnaire was to obtain expert comments on the highlighted regions' alignment without altering the model architecture to preserve the performance of the model. The combination of the EfiicientNetB7 model as a feature extractor with an SVM activation function outperformed and achieved the accuracy and the area under the curve (AUC) of 98.95% and 0.9886, respectively. This research contributes to the ongoing efforts to make deep learning models for breast cancer diagnosis more transparent and trustworthy.

Keywords: Breast cancer diagnosis; EfficientNet; GradCAM; Histopathology visualization; Heatmap, Interpretable Model

1. Introduction

With the use of interpretable machine learning models, clinicians would be better able to comprehend the rationale behind the model's recommendations and make more educated treatment decisions when results are presented understandably. For example, the Breast Cancer Treatment Recommender System uses patient data to provide suggestions for individualized treatment plans based on the patient's medical history, lifestyle, and cancer subtype, among other things. Interpretable machine learning models play a pivotal role in aiding clinicians' treatment decisions for breast cancer by[1]–[4] :

- **Increased comprehension and trust:** Clinicians can comprehend the reasoning behind a specific treatment prescription made by machine learning models that are transparent in their decision-making processes. This openness increases confidence in the AI systems.
- **Tailored Care Programs:** These models offer individualized treatment approaches based on factors

specific to each patient using, which clinicians can increase the efficacy of treatments for breast cancer.

- **Early Detection and Diagnosis:** Mammograms and other medical imaging data can be precisely analyzed by machine learning models. They can support early diagnosis and detection of breast cancer for more effective therapy and improved patient outcomes.
- **Reduced Error and Variability:** Interpretable models contribute to a decrease in diagnostic errors and treatment decision variations by offering data-driven insights.
- **Integration into the clinical workflow:** Healthcare workers can more easily access and use AI-driven advice in addition to their clinical experience when interpretable models are incorporated into clinical workflows.
- **Ethical and Legal Considerations:** Interpretable models guarantee that recommendations generated by AI follow moral and legal requirements. In the healthcare industry, patient safety and data privacy are of utmost importance.

The visualization technique developed in this study not only addresses concerns about their black-box nature but also offers valuable insights for healthcare professionals in diagnostic decision-making. Notably, the model's accuracy remains the same despite the unchanged architecture, promising significant contributions to enhancing trust in automated systems for medical applications, notably in breast cancer diagnosis.

The paper is organized as follows: Section 2 provides a comprehensive literature review, delving into earlier research. Details regarding the materials and procedures employed are presented in Section 3. Section 4 presents the results and subsequent discussion. Finally, Section 5 encapsulates the conclusion and outlines future directions for the study.

2. Literature Review

Decision Support Systems(DSS) that use machine learning are becoming more and more common. The increasing use of these systems has sped up the transition to an increasingly computational society, increasing the likelihood that judgments made using algorithmic intelligence will have a big societal impact. However, the majority of these precise DSSs are still complicated "black boxes," meaning that even specialists are unable to completely comprehend the reasoning behind the systems' predictions due to their internal workings and concealed logic from the user.[5]

Explainability and interpretability are closely associated concepts. However, it is also noted that the term "interpretable" is more commonly used in the machine learning community than "explainable."[5] Machine learning models like Support Vector Machine, Naïve Bayes, K-nearest neighbors, AdaBoost, and LightGBM have been applied for breast cancer prediction[6], [7]. These models are designed with built-in transparency and interpretability, which enables doctors and patients to comprehend how the model generates its diagnoses. Interpretability methodologies like SHAP (Shapley Additive Explanations) [8] are particularly instrumental in elucidating these models' decisions, offering crucial insights for medical professionals to make well-informed decisions regarding patient care and treatment plans [9], [10], [11]. These models heavily rely on human-crafted features engineered by domain experts to yield accurate predictions.

Conversely, deep learning models possess the capability to automatically extract features from raw data, eliminating the necessity for manual feature engineering. It enables us to comprehend precisely what a model is learning, what further information the model has to provide, and the reasoning behind its decisions[12], [13]. However, despite their ability to self-learn features, deep learning models lack inherent interpretability features, making it challenging to comprehend their decision-making process. The advantages and distinctions between explainable and standard machine learning models are depicted in Figure 1. Explainable models help us understand how the model works and what factors are influencing its decisions. This can help build trust, causality, and informativeness in the model and ensure that the model is not biased or discriminatory [5], [15], [16], [17]. Explainable models can help us understand how these decisions were made and who is responsible for them. Some industries and regulatory bodies require that machine learning models be explainable to comply with regulations and guidelines [4], [17], [18]. Figure 2 illustrates the accuracy and interpretability of several machine learning algorithms [14]. In essence, conventional ML involves interpretable steps, whereas CNNs prioritize performance over explainability, making their decision-making process less transparent.[19].LIME (Local Interpretable Model-agnostic Explanations) generates local explanations for individual predictions by perturbing the input data and observing how the model's predictions change. This can provide insight into the specific factors that are influencing the model's decision for a particular case [20]–[22].

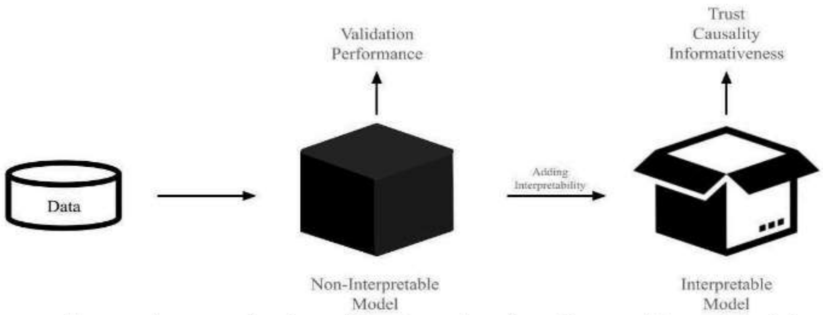


Figure 1: Comparing standard machine learning algorithms with explainable ones

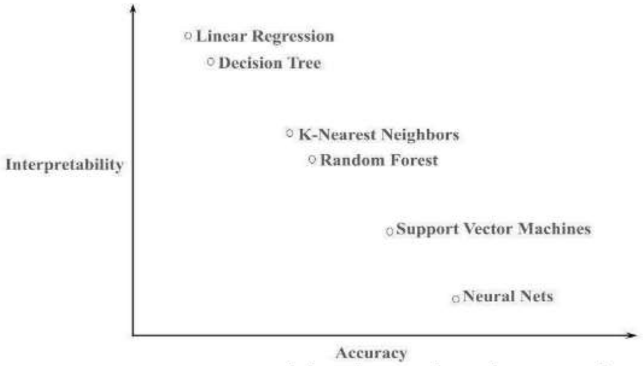


Figure 2: Accuracy versus interpretability for machine learning algorithms [14].

The following describes the specifics of the primary methods for enhancing interpretability in deep learning models:

o **Global Average Pooling (GAP)**: It is a technique used to reduce the spatial dimensions of a 3D tensor by taking the average of all values in each feature map. It condenses convolutional features for better localization without fully connected layers. In the deep learning model CAM and GAP layer after the last convolution layer and removing the fully connected layer in the model [23] generate heatmaps for each output class.

o **Class Activation Map (CAM)**: It is a technique used to visualize the areas of an image that contribute the most to a specific class prediction in a convolutional neural network. It does this by generating a heatmap that highlights the important regions of the input image. CAM utilizes a GAP layer post-final convolutional layers, highlighting crucial image regions for specific predictions.

o **Grad-CAM (Gradient-weighted Class Activation Mapping):** It produces the activations without changing the model's architecture [24] by computing the first-order gradients concerning the final convolutional layer.These gradients highlight what the model focuses on when making predictions. Grad-CAM is an improvement over CAM, offering broader applicability without architectural restrictions [24], [25] [26]. Grad-CAM reveals which areas in the last convolutional layer's feature maps are activated, aiding in understanding feature importance [20], [21]. Augmented Grad-CAM [27] provides a high-resolution visual explanation of deep neural networks by increasing the resolution of heatmaps through augmentation. [28]

o **Grad Cam++:** Grad Cam++ uses the first-order gradients of the output class score concerning the feature maps of the last convolutional layer to generate more accurate and sharper heat maps. It refines Grad-CAM, utilizing methods like SmoothGrad or Expected Grad-CAM, improving visualization accuracy [28]. Smooth-grad++ [29] introduces noise in the images and then computes the gradients and heatmaps. Grad Cam++ [30] produces more sharp heatmaps and increases the localization accuracy. It computes higher-order derivatives. Table 1 lists a few deep learning models along with datasets and performance metrics that have been used to diagnose breast cancer in the literature. Table 2 provides a summary of previous work in the domain of interpretability of deep learning models for cancer diagnosis using Grad-CAM.

Table 1: current standards Models of deep learning applied to the detection of breast cancer

Reference	**Name of Model**	**Dataset**	**Task**	**Performance Metric**
[31]	InceptionV4	ICIAR-2018	Binary Classification	Accuracy:93.7%
[32]	DenseNet 121with SENet	BeakHis	Binary Classification	PRR:89.5, IRR:89.1
[33]	MobileNet and EfficientNet-B3	TCGA	Segmentation	Sensitivity99%
[34]	EfficientNet-B3	RPCam	Binary Classification	Accuracy: 97.9%

[35]	EfficientNet-B6	RPCam	Binary Classification	Accuracy: 97.94%
[36]	EfficientNet-B2	ICIAR-2018	Multi-class Classification	Accuracy: 98.33%

Table 2: Summary of previous works in literature utilized Grad-CAM for making deep learning models interpretable in biomedical imaging

Reference	Name of Model	Types of Images	Accuracy/ AUC
[37]	VGGNet	Mammographic Image Analysis Society (MIAS) and Digital Database for Screening Mammography (DDSM)	92%
[38]	VGG16 ResNet50 Alex_Net, and MobileNet	Br35H::Brain Tumor Detection 2020	97.83% 99.67% 99.3% 98.5%
[39]	CNN model from scratch	A private dataset consisting of X-ray images of Covid 19	98%
[40]	3D CNN model	LUNA 16 Dataset{ X-ray images of Lung cancer}	0.97
[41]	CNN	CT images For neck and head cancer	0.92
[42]	DenseNet-169 EfficientNet-B5	A private dataset consisting of Mammography	0.952 ± 0.005 0.954 ± 0.020
[43]	squeeze Net	Histopathology Breast cancer Image dataset	90.3%

3. Method and Materials

This section provides an overview of the dataset utilized for training and validation, offering a concise preview of its key parameters. Subsequently, the tools employed in the study are itemized. The methodology is then explained, detailing the workflow implemented throughout the research process. Additionally, detailed insights are provided regarding the application of EfficientNet and GradCAM.

3.1 *Dataset*

Histopathology Cancer Detection (HCD)

The Histopathology Cancer Detection dataset was created as part of a Kaggle competition to create algorithms that could recognize metastatic cancer in lymph node sections' histological pictures. Many histopathologic pictures of lymph nodes, each categorized as positive (cancerous) or negative, are included in the dataset (non-cancerous). The dataset comprises histopathologic images that are employed in the process of identifying cancer in tissue from lymph nodes. It focuses on identifying metastatic tissue in these photos specifically. 327,680 color photos, each measuring 96 x 96 pixels, make up the dataset. This dataset has been used as the foundation for cancer detection-related Kaggle competitions. The distribution of labels in the dataset and sample images are shown in Figure 3 and Figure 4, respectively.

Web Link: https://www.kaggle.com/c/histopathologic-cancer-detection/data

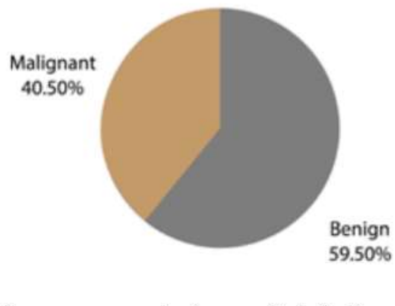


Figure 3: sample image distribution in the dataset

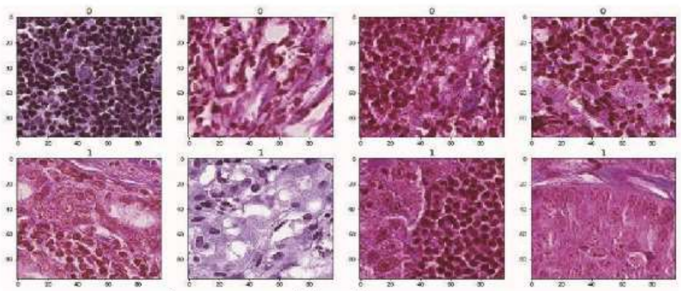


Figure 4: Sample images of the dataset

3.2 *Tools*

Python is used to implement the models utilizing the TensorFlow and Keras frameworks as the backend. The Keras library was used to import the EfficientNet models (B0-B7).The Kaggle kernel on the GPU virtual machine was used for all of the experiments, with the following specifications:

GPU: Nvidia Tesla P100; performance: 9.3 TFLOPS; GPU memory: 16GB; GPU memory clock: 1.32 GHz

3.3 Workflow

Figure 5 shows the workflow of the suggested methodology utilized for expert validation with binary breast cancer classification and heatmap generation. The accuracy, AUC, and recall are used to assess the performance of the models. The next step after binary classification is to use the Grad-CAM algorithm to create heat maps. In addition to displaying the label of the image, this attempts to provide visual explanations by highlighting the regions that are involved in classifications. A questionnaire was administered to pathologists, presenting snippets of the output generated by the model for validation. The opinions of the pathologists are included after the provided snippets shown in the original image.

3.4 EfficientNet

EfficientNet is a group of robust convolutional neural network (CNN) architectures, namely EfficientNetB0-EfficientNetB7, renowned for their efficiency and accuracy in image classification tasks. These models boast a sophisticated architecture that balances depth, width, and resolution to optimize performance. EfficientNet architecture consists of seven distinct blocks and involves a hierarchy of layers within each block, incorporating intricate arrangements of depth-wise separable convolutions, normalization layers, and shortcut connections, contributing to its impressive efficiency and accuracy. In the experiment, EfficeientNet architectures B0-B7 were used with transfer learning for feature extraction, and then SVM was applied for binary classification of histopathology images from the HCD dataset. Additionally, data augmentation was applied to increase the images in the dataset. Training and testing sets were divided into an 80:20 ratio in the dataset. The 96x96x3 histopathological image is sent into the model's input layer. Concatenation of "GlobalmaxPooling", "GlobalAverragePooling", and "Flatten" layers followed by dropout layer, and finally, "sigmoid" layers or SVM(Support Vector Machine) are added in case of fine-tuning and feature extractor respectively after the "efficientNet-b7"(base model). An Adam optimizer with a learning rate of 0.001 and a binary cross entropy loss function was used to compile the model.

3.4.1 Grad-CAM

Grad-CAM is an inference tool that generates graphical information by extracting gradients from model convolutional layers. These gradients identify key areas in input photos and represent high-level visual patterns. Grad-CAM uses the spatial information that convolutional layers store to produce heatmaps that emphasize the areas of the image that influence model selections. This gives the deep learning model's judgments a visual justification. As a visualization method for comprehending CNNs and their decision-making processes, the Grad-CAM has several benefits, including interpretability, no architecture modification, localization, and applicability to different tasks. The experiment's algorithm is provided in Algorithm 1 for understanding.

Algorithm 1. Working Procedure of *make_gradcam_heatmap*

Input: img, model, efn_model,conv_layer_name

Output: heatmap of the histopathology image and the label of the image

```
Begin
    img_arrày→Prepares the input image for processing
    conv_layer→efn_model.get_layer(conv_layer_name)
    conv_layer_model→ Model(efn_model.inputs,conv_layer.output)
    classifier_model→Model(conv_layer_model,output,efn_model.output)
    classifier_model→Model(classifier_model.input,model.layers)
    conv_layer_output→ classifier_model(img-arr)
    pred→ classifier_model(conv_layer_output)
    grads→ GradientTape.gradient( pred , conv_layer_output)
    conv_layer_output→conv_layer_output *grads
    heatmap→mean(con_layer_output)
    label→[1 if pred>=0.5 else 0]
End
```

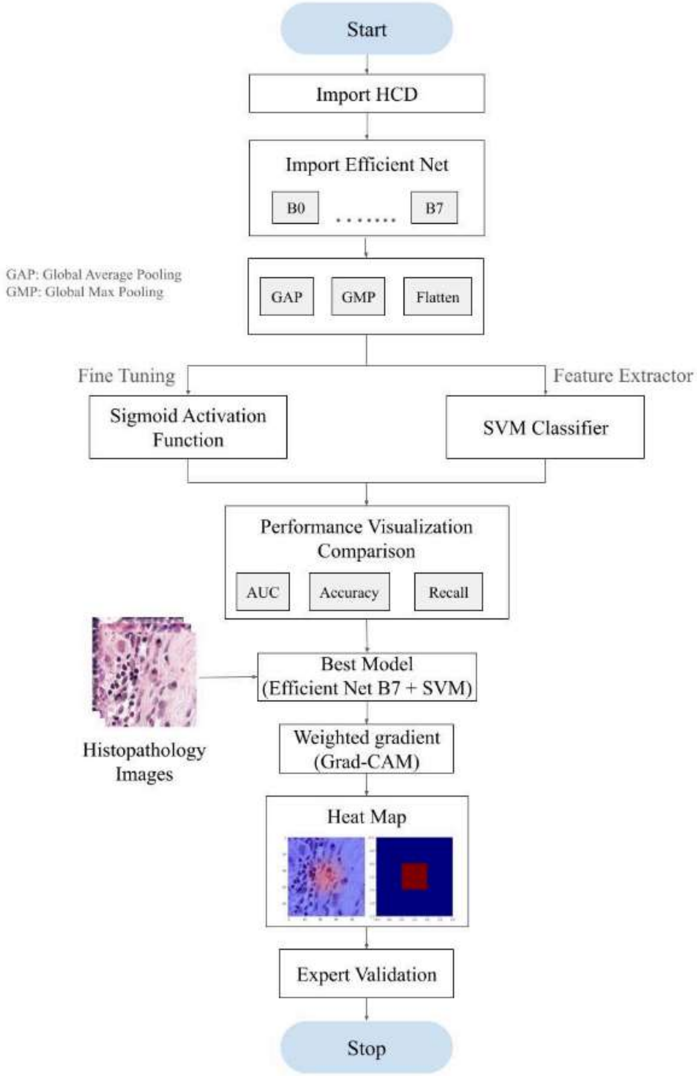


Figure 5: Workflow of the proposed methodology used for generating heatmaps and expert validation with binary breast cancer classification

4 Results and Discussion

This section is divided into two subsections. The first subsection presents the results achieved for binary classification of histopathology images by EfficienNet models as fine-tuning and as a feature extractor. The second subsection presents the results obtained to utilize the best model for binary classification with Grad-CAM for visualizing the output of a specified block of the model, generating the heatmaps and the output achieved by superimposing the generated heatmap.

4.1 *Binary Classification of Histopathology Images*

The performance metrics of the proposed models as feature extractor and fine-tuning are shown in Table 3 and Table 4, respectively.

Table 3: Performance matrices of EfficientNet as feature extractor for binary classification

Model	**Validation Accuracy**	**AUC**	**Precision**		**Recall**		**F1- score**	
			Benign	**Malignant**	**Benign**	**Malignant**	**Benign**	**Malignant**
B0 + SVM	0.9835	0.9829	0.9865	0.979	0.9858	0.9801	0.9861	0.9796
B1 + SVM	0.9856	0.9852	0.9885	0.9815	0.9874	0.983	0.9879	0.9822
B2 + SVM	0.9871	0.9866	0.9889	0.9845	0.9895	0.9837	0.9892	0.9841
B3 + SVM	0.9876	0.9872	0.9897	0.9846	0.9896	0.9848	0.9896	0.9847
B4 + SMV	0.9889	0.9887	0.9916	0.985	0.9898	0.9876	0.9907	0.9863

B5 + SVM	0.9887	0.9884	0.9911	0.9852	0.99	0.9868	0.9905	0.986
B6 + SVM	0.9895	0.9892	0.9917	0.9862	0.9906	0.9879	0.9912	0.987
B7 + SVM	0.9895	0.9891	0.9913	0.9868	0.991	0.9872	0.9912	0.987

Table 4: Performance matrices of EfficientNet as fine-tuning for binary classification

Model	Validation Accuracy	AUC	Precision		Recall		F1- score	
			Benign	Malignant	Benign	Malignant	Benign	Malignant
B0	0.9818	0.9800	0.9803	0.984	0.9893	0.9707	0.9848	0.9773
B1	0.9837	0.9831	0.9866	0.9793	0.9859	0.9803	0.9863	0.9798
B2	0.9855	0.9853	0.9894	0.98	0.9863	0.9844	0.9878	0.9822
B3	0.9859	0.9858	0.9902	0.9795	0.986	0.9857	0.9881	0.9826
B4	0.9881	0.9879	0.991	0.9837	0.9889	0.9868	0.99	0.9853
B5	0.987	0.9868	0.9903	0.9822	0.9879	0.9857	0.9891	0.9839
B6	0.9888	0.9886	0.9894	0.9879	0.9918	0.9844	0.9906	0.9861
B7	0.9889	0.9886	0.9911	0.9856	0.9902	0.987	0.9907	0.9863

4.2 Visualization of the layers of the model and heatmap generation

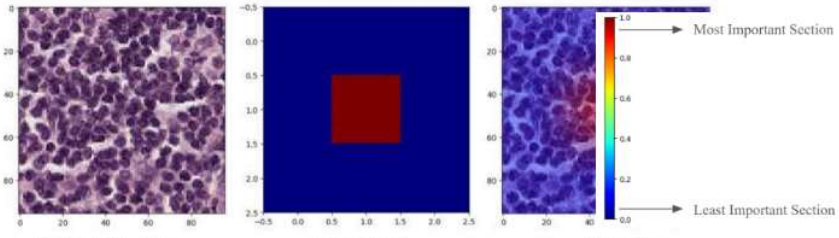
In this section, histopathology images, paired with their respective heatmaps and overlays, are generated by the model. The expert validation scores are included in the notes following the snippet. Four sets of images are included. Figures 6 and 9 include heat maps of Histopathology Images classified correctly as Benign and Malignant, respectively. Figures 8 and 11 include heat maps of Histopathology images classified incorrectly as Malignant and Benign, respectively. Figure 7 and Figure 10 dissect the output of the convolution layer across the seven blocks of the EfficinetNetB7 model when it correctly classifies Benign and Malignant tissue from the histopathology image, respectively.

4.3 Expert Validation

Three columns consisting of the following image were presented to expert and experienced pathologists:

1. Original Image: It is a copy of the original biopsy slide.
2. Overlay: It is the biopsy slide with the important section marked in red and unimportant parts marked in blue.
3. Heat Map: It is the heat map's importance across the biopsy slide.

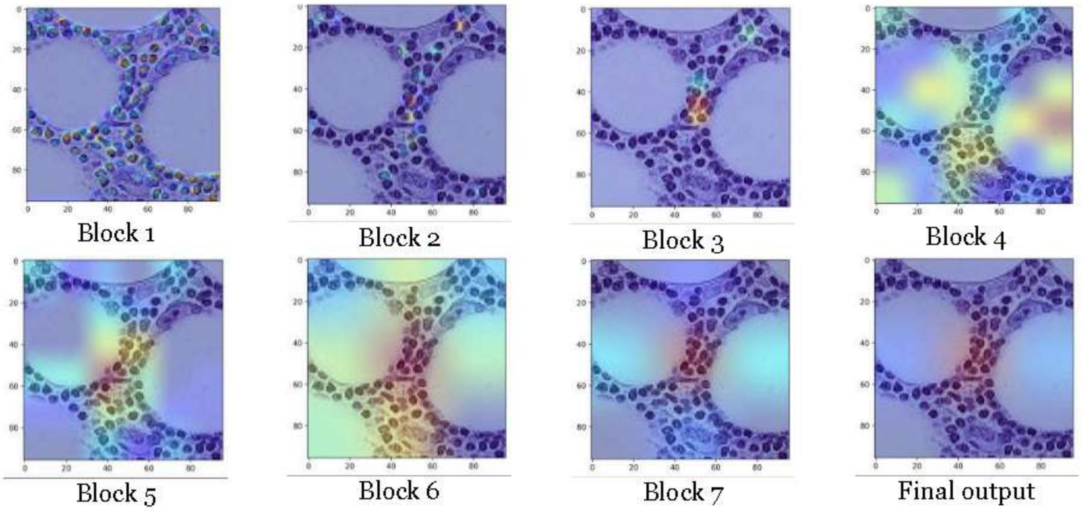
The expert is asked to judge the performance of the model by seeing the highlighted region marked by the model.



Original Image Heat Map Overlay

Figure 6: Original Image and Model Output where Benign slides classified as Benign

Case 1: The biopsy slide shown in Figure 6 showcases **Benign** tissue samples accurately classified by our model, with the highlighted areas of interest depicted in red within the accompanying heatmaps. The model's precision is validated by **four out of four** pathologists in the survey, who independently identified and concurred with the highlighted regions.



Block 1

Block 2

Block 3

Block 4

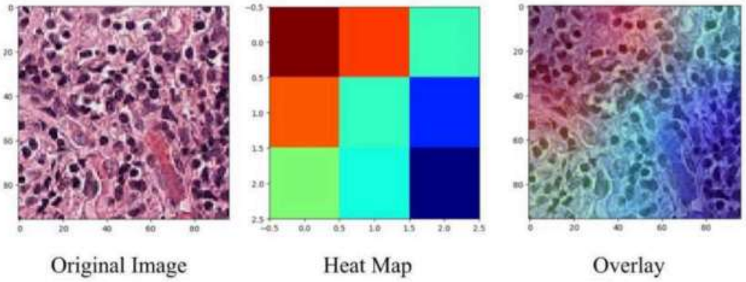
Block 5

Block 6

Block 7

Final output

Figure 7:Convolution layer dissection across the seven blocks of the EfficinetNetB7 model with Grad-CAM when it correctly classifies Benign histopathology image



Original Image

Heat Map

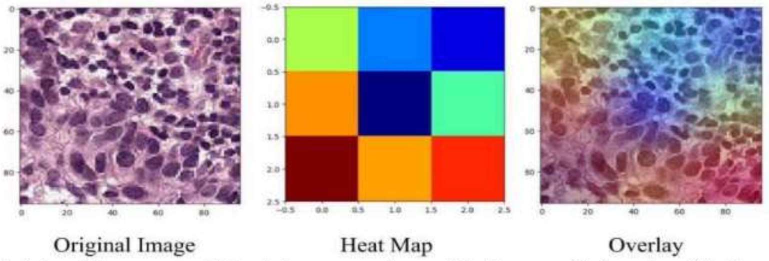
Overlay

Figure 8: Original Image and Models output where Benign slides classified as Malignant

Case 2: The biopsy slide in Figure 8 depicts benign tissue samples misclassified as malignant by the model, with the heatmap revealing the specific section leading to this erroneous decision. Notably, **three out of four** pathologists in our survey acknowledge the model's confusion and advocate for additional clinical context to enhance accuracy.

Case 3 *The* biopsy slide presented in Figure 9 features malignant tissue accurately classified by our model, with the highlighted areas of interest depicted in red within the accompanying heatmaps. It is noteworthy that 3/4 pathologists in our survey concur with the model's highlighted regions. This variance in expert opinions underscores the complexity of pathology interpretation and emphasizes the potential complementarity between machine learning models and human expertise in refining diagnostic accuracy.

Case 4: Figure 11 presents the biopsy slide exhibiting malignant tissue incorrectly classified by the model, with the highlighted areas of interest shown in red on the accompanying heatmaps. Notably, three out of four pathologists in our survey acknowledge the model's confusion and advocate for additional clinical context to enhance accuracy.

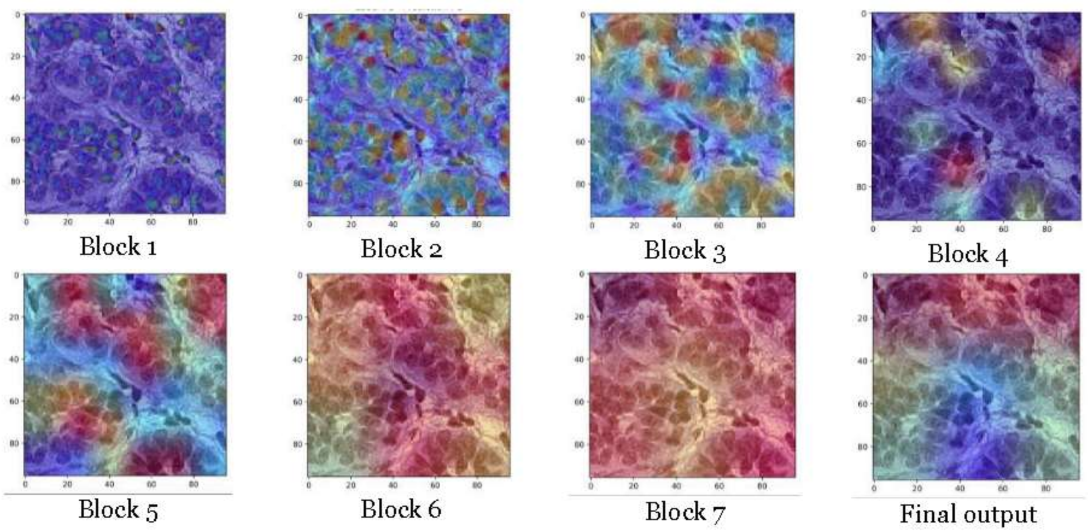


Original Image

Heat Map

Overlay

Figure 9: Original Image and Model output where Malignant slides classified as Malignant



Block 1

Block 2

Block 3

Block 4

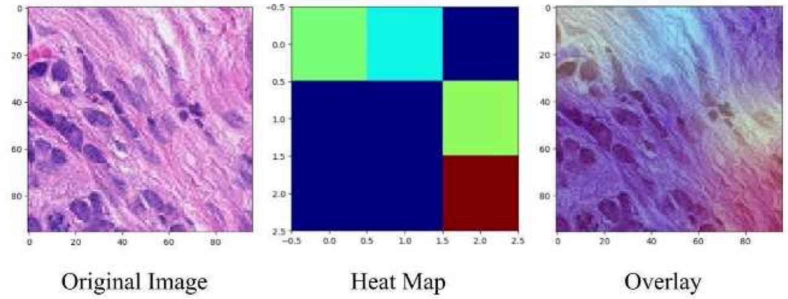
Block 5

Block 6

Block 7

Final output

Figure 10:Convolution layer dissection across the seven blocks of the EfficinetNetB7 model with Grad-CAM when it correctly classifies Malignant histopathology image



Original Image

Heat Map

Overlay

Figure 11: Original Image and Models Output where Malignant slides classified as Benign

5 Conclusion

In conclusion, the timely and accurate diagnosis of breast cancer is paramount for improving survival rates, emphasizing the significance of advanced diagnostic tools. In this study, the utilization of EfficientNet B7 achieved an impressive 99.89% accuracy, laying a robust foundation for reliable breast cancer diagnosis. Notably, this accuracy surpasses that reported in related studies highlighted in the literature review. The alteration of the model architecture to incorporate visualization through heatmaps provided a crucial step toward enhancing interpretability in deep learning-based diagnostics. The introduction of explainable deep learning, manifested through heatmaps, offers healthcare professionals valuable insights into the decision-making process of the model. However, it is essential to acknowledge the limitations of this study. The trained model's lack of diversity hinders its broad applicability to various histopathology images, and the validation, though conducted with pathologists, was limited in number and lacked unanimity. These limitations underscore the need for future developments in creating more versatile models and expanding validation efforts to ensure the robustness and generalizability of the proposed approach. Despite these challenges, this research lays a promising foundation for further advancements in transparent and trustworthy deep-learning models for breast cancer diagnosis.

References

1. F. Di Martino and F. Delmastro, "Explainable AI for clinical and remote health applications: a survey on tabular and time series data," *Artif Intell Rev*, vol. 56, no. 6, pp. 5261–5315, Jun. 2023, doi: 10.1007/s10462-022-10304-3.
2. T. Khater, S. Ansari, S. Mahmoud, A. Hussain, and H. Tawfik, "Skin cancer classification using explainable artificial intelligence on pre-extracted image features," *Intelligent Systems with Applications*, vol. 20, Nov. 2023, doi: 10.1016/j.iswa.2023.200275.
3. T. Suresh, T. A. Assegie, S. Ganesan, R. L. Tulasi, R. Mothukuri, and A. O. Salau, "Explainable extreme boosting model for breast cancer diagnosis," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, pp. 5764–5769, Oct. 2023, doi: 10.11591/ijece.v13i5.pp5764-5769.

4. "Improve explainability of ML models to meet regulatory requirements - Amazon Science." Accessed: Dec. 24, 2023. [Online]. Available: https://www.amazon.science/latest-news/remars-revisited-improve-explainability-of-ml-models-to-meet-regulatory-requirements
5. D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics (Switzerland)*, vol. 8, no. 8, Aug. 2019, doi: 10.3390/ELECTRONICS8080832.
6. K. M. Mohi Uddin, N. Biswas, S. T. Rikta, S. K. Dey, and A. Qazi, "XML-LightGBMDroid: A self-driven interactive mobile application utilizing explainable machine learning for breast cancer diagnosis," *Engineering Reports*, vol. 5, no. 11, Nov. 2023, doi: 10.1002/ENG2.12666.
7. K. Mohammad, M. Uddin, N. Biswas, S. Tasmin Rikta, and S. Kumar Dey, "Machine learning-based diagnosis of breast cancer utilizing feature optimization technique," 2023, doi: 10.1016/j.cmpbup.2023.100098.
8. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 1135–1144, Aug. 2016, doi: 10.1145/2939672.2939778.
9. "Explainable Deep Learning in Breast Cancer Prediction | by Yu Huang, M.D., M.S. in CS | Towards Data Science." Accessed: Dec. 24, 2023. [Online]. Available: https://towardsdatascience.com/explainable-deep-learning-in-breast-cancer-prediction-ae36c638d2a4
10. T. Brito-Sarracino, M. Rocha Dos Santos, E. Freire Antunes, I. Batista De Andrade Santos, J. Coelho Kasmanas, and A. C. Ponce De Leon Ferreira De Carvalho, "Explainable machine learning for breast cancer diagnosis," *Proceedings - 2019 Brazilian Conference on Intelligent Systems, BRACIS 2019*, pp. 681–686, Oct. 2019, doi: 10.1109/BRACIS.2019.00124.
11. J. Vrdoljak *et al.*, "Applying Explainable Machine Learning Models for Detection of Breast Cancer Lymph Node Metastasis in Patients Eligible for Neoadjuvant Treatment," *Cancers (Basel)*, vol. 15, no. 3, Feb. 2023, doi: 10.3390/cancers15030634.
12. U. Johansson, C. Sönströd, U. Norinder, and H. Boström, "Trade-off between accuracy and interpretability for predictive in silico modeling," *Future Med Chem*, vol. 3, no. 6, pp. 647–663, 2011, doi: 10.4155/FMC.11.23.
13. A. Chatzimparmpas, R. M. Martins, and A. Kerren, "A survey of surveys on the use of visualization for interpreting machine learning models," *Article Information Visualization*, vol. 2020, no. 3, pp. 207–233, doi: 10.1177/1473871620904671.
14. M. A. Gulum, C. M. Trombley, and M. Kantardzic, "A Review of Explainable Deep Learning Cancer Detection Models in Medical Imaging," *Applied Sciences 2021, Vol. 11, Page 4573*, vol. 11, no. 10, p. 4573, May 2021, doi: 10.3390/APP11104573.
15. S. C. Lu, C. L. Swisher, C. Chung, D. Jaffray, and C. Sidey-Gibbons, "On the importance of interpretable machine learning predictions to inform clinical decision making in oncology," *Front Oncol*, vol. 13, p. 1129380, Feb. 2023, doi: 10.3389/FONC.2023.1129380/BIBTEX.
16. E. Onose, "Explainability and Auditability in ML: Definitions, Techniques, and Tools," neptune.ai. [Online]. Available: https://neptune.ai/blog/explainability-auditability-ml-definitions-techniques-tools
17. "Explainable AI: Getting it right in business | McKinsey." Accessed: Dec. 24, 2023. [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-businesses-need-explainable-ai-and-how-to-deliver-it
18. A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.
19. "4. Explainability for Image Data - Explainable AI for Practitioners [Book]." Accessed: Dec. 24, 2023. [Online]. Available: https://www.oreilly.com/library/view/explainable-ai-for/9781098119126/ch04.html
20. D. Wu and J. Zhao, "Understand how CNN diagnoses faults with Grad-CAM," *Computer Aided Chemical Engineering*, vol. 49, pp. 1537–1542, Jan. 2022, doi: 10.1016/B978-0-323-85159-6.50256-6.
21. I. E. Nielsen *et al.*, "Robust Explainability: A tutorial on gradient-based attribution methods for deep neural networks," vol. 39, no. 4, pp. 73–84, 2022, doi: 10.1109/MSP.2022.3142719.
22. Q. Zhang, Y. N. Wu, and S. C. Zhu, "Interpretable Convolutional Neural Networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8827–8836, Oct. 2017, doi: 10.1109/CVPR.2018.00920.
23. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization".
24. "Grad-CAM: Visualize class activation maps with Keras, TensorFlow, and Deep Learning - PyImageSearch." Accessed: Dec. 12, 2023. [Online]. Available: https://pyimagesearch.com/2020/03/09/grad-cam-visualize-class-activation-maps-with-keras-tensorflow-and-deep-learning/
25. "Understand your Algorithm with Grad-CAM | by Daniel Reiff | Towards Data Science." Accessed: Dec. 12, 2023. [Online]. Available: https://towardsdatascience.com/understand-your-algorithm-with-grad-cam-d3b62fce353

26. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization".
27. P. Morbidelli, D. Carrera, B. Rossi, P. Fragneto, and G. Boracchi, "AUGMENTED GRAD-CAM: HEAT-MAPS SUPER RESOLUTION THROUGH AUGMENTATION".
28. R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, "Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs," Aug. 2020, [Online]. Available: http://arxiv.org/abs/2008.02312
29. D. Omeiza, S. Speakman, C. Cintas, and K. Weldemariam, "Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models".
30. A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks".
31. M. I. Sarker, H. Kim, D. Tarasov, and D. Akhmetzanov, "Inception Architecture and Residual Connections in Classification of Breast Cancer Histology Images Inception Architecture and Residual Connections in Classification of Breast Cancer Histology Images," no. December 2019, 2020.
32. X. Li, X. Shen, Y. Zhou, X. Wang, and T. Q. Li, "Classification of breast cancer histopathological images using interleaved DenseNet with SENet (IDSNet)," *PLoS One*, vol. 15, no. 5, pp. 1–13, 2020, doi: 10.1371/journal.pone.0232127.
33. G. Gagan *et al.*, "Unmet clinical need: Developing prognostic biomarkers and precision medicine to forecast early tumor relapse, detect chemo-resistance and improve overall survival in high-risk breast cancer," *Ann. Breast Cancer Ther.*, vol. 4, no. 1, pp. 48–57, May 2020, doi: 10.36959/739/525.
34. J. Wang, Q. Liu, H. Xie, Z. Yang, and H. Zhou, "Boosted EfficientNet: Detection of Lymph Node Metastases in Breast Cancer Using Convolutional Neural Network," Oct. 2020.
35. Y. Sun, F. A. Binti Hamzah, and B. Mochizuki, "Optimized Light-Weight Convolutional Neural Networks for Histopathologic Cancer Detection," in *LifeTech 2020 - 2020 IEEE 2nd Global Conference on Life Sciences and Technologies*, Institute of Electrical and Electronics Engineers Inc., Mar. 2020, pp. 11–14. doi: 10.1109/LifeTech48969.2020.1570619224.
36. C. Munien and S. Viriri, "Classification of Hematoxylin and Eosin-Stained Breast Cancer Histology Microscopy Images Using Transfer Learning with EfficientNets," *Comput Intell Neurosci*, vol. 2021, 2021, doi: 10.1155/2021/5580914.
37. H. Wang *et al.*, "Breast Mass Detection in Digital Mammogram Based on Gestalt Psychology," *J Healthc Eng*, vol. 2018, 2018, doi: 10.1155/2018/4015613.
38. F. Mercaldo, L. Brunese, F. Martinelli, A. Santone, and M. Cesarelli, "Explainable Convolutional Neural Networks for Brain Cancer Detection and Localisation," *Sensors*, vol. 23, no. 17, Sep. 2023, doi: 10.3390/s23177614.
39. C. V. Aravinda, M. Lin, K. R. Udaya Kumar Reddy, and G. A. Prabhu, "A demystifying convolutional neural networks using Grad-CAM for prediction of coronavirus disease (COVID-19) on X-ray images," *Data Science for COVID-19*, p. 429, Jan. 2021, doi: 10.1016/B978-0-12-824536-1.00037-X.
40. E. Stephen, N. Joshua, D. Bhattacharyya, M. Chakkravarthy, and Y.-C. Byun, "3D CNN with Visual Insights for Early Detection of Lung Cancer Using Gradient-Weighted Class Activation," 2021, doi: 10.1155/2021/6695518.
41. A. Diamant, A. Chatterjee, M. Vallières, G. Shenouda, and J. Seuntjens, "Deep learning in head & neck cancer outcome prediction," *Sci Rep*, vol. 9, no. 1, Dec. 2019, doi: 10.1038/s41598-019-39206-1.
42. Y. J. Suh, J. Jung, and B. J. Cho, "Automated breast cancer detection in digital mammograms of various densities via deep learning," *J Pers Med*, vol. 10, no. 4, pp. 1–11, Nov. 2020, doi: 10.3390/jpm10040211.
43. S. Chaudhury, K. Sau, M. A. Khan, and M. Shabaz, "Deep transfer learning for IDC breast cancer detection using fast AI technique and Sqeezenet architecture," *Mathematical Biosciences and Engineering*, vol. 20, no. 6. American Institute of Mathematical Sciences, pp. 10404–10427, 2023. doi: 10.3934/mbe.2023457.