



## Diagnosis and Interpretation of Breast Cancer Using Explainable Artificial Intelligence

Francis A. U. Imouokhome, Osehi Grace Ehimiyein and Fidelis Odinma Chete\*

Department of Computer Science, University of Benin, Benin, City

### Article Info

#### Keywords:

Breast Cancer, Resnet50, Benign, Malignant, eXplainable Artificial Intelligence, Deep Learning

Received 11 April 2023

Revised 26 April 2023

Accepted 27 April 2023

Available online 07 June 2023

<https://doi.org/10.5281/zenodo.8014197>

ISSN-2682-5821/© 2023 NIPES Pub. All rights reserved.

### Abstract

Breast cancer is one of the leading causes of death among women and timely intervention is the key to curb it. This has necessitated Information Technology (IT) researchers and professionals to continually create models that can help in early detection of breast cancer, the area of interpretation has, however, not been explored. This has motivated this research to visually interpret breast cancer diagnosis to Pathologists and even layman who wishes to know. In this research the BreakHis dataset from Kaggle Challenge was used. A ResNet50 model (adopted in this research) was trained, using deep learning in order to classify the breast tumor as either malignant or benign. The result obtained from testing the model was 96.84% which outperformed results achieved by other researchers who used the same deep learning methodology. The classification of breast cancer diagnosis from histopathological images were later interpreted using eXplainable Artificial Intelligence (AI) techniques like Integrated Gradient (IG), GradientShap (GS) and Occlusion, which gave reasons why a particular histopathological image was considered as Benign or Malignant. Comparing these three techniques, Occlusion was found to have more predictive results based on visualization and time of execution. This research did not only classify histopathological images as either benign or malignant but also gave reasons for its results unlike other earlier studies.

### 1.0. Introduction

Cancer is the universal name given to an abnormal cell growth in the human body. It is one of the prominent causes of death worldwide. In all types of cancerous cells, numerous body tissues divide and spread wildly around cells. Cancer can start almost anywhere in the human body. Human tissues grow and divide to form new tissues when needed in the body. In a normal situation, as cells become older, they die and are replaced with new ones but when cancer occurs, this orderly process changes; older cells which are supposed to die and be replaced, survive, and new ones which are not needed are formed [1] These extra cells that are formed divide continuously and form abnormal growths which are referred to as tumors. Numerous types of cancer form solid tumors, which are composed of cell masses. Tumors can either be benign or malignant [2]. A benign tumor is non-cancerous, it grows slowly but cannot spread to other parts of the body; if operated on and the tumor does not generally return. Malignant tumors are cancerous and the cells grow and spread to other parts of the body uncontrollably. According to evaluations from the International Agency for Research on Cancers, 14.1 million new cancer cases were diagnosed and 8.2 million people died from cancer worldwide in 2012 [3]. Without treatment, cancer can cause serious health problems and even loss

of lives. Early detection is the key to reducing the mortality rate. Cancer can occur in any part of the human body, liver, ovary, pancreas, lungs, breast and so on. Cancer that occurs around the breast region of humans (male or female) is referred to as breast Cancer. It can either be an epithelial tumor or non-epithelial tumor. Epithelial tumor arises from the milk producing region or the draining duct of the breast and non epithelial tumor occur from the soft tissues of the breast [4]. Various techniques which have been employed to detect breast cancer include Mammography, Magnetic Resonance Imaging (MRI) Scans, Computed Tomography (CT) Scans, Ultrasound, Nuclear Imaging, amongst others. Among the various cancers known, breast cancer happens to occur more in women. According to [5], breast cancer has the second highest mortality rate after Lung and Bronchial cancer, and about 30% of newly diagnosed cases are of breast cancer. Early diagnosis of these cancer cases is one of the steps to curbing the menace.

Several years ago, several breast cancer diagnostic models were proposed by some scholars [6]-[7], using Machine Learning, but it lacked eXplainability as to why the results of diagnosis is what it is. This is particularly true for a “blackbox” approach like deep learning, which has created a lot of chaos amongst Pathologists. However, in medical systems, black boxes are usually not well-appreciated by physicians since they prefer to understand how the system produces recommendations [8]. Thus, an eXplainable model is needed to solve this challenge.

It has been observed from the previous researches that physicians find it difficult to give a vivid reason why certain decisions are made by machines during diagnosis, due to its opaque nature. This has had adverse impact on decision making. To bridge this gap, an insight into the internal operation of each model layer must be visually exposed or explained. This would lead to an effective interpretation of the neural network predictions by visualizing its decision operations, which in this case is diagnosing breast cancer disease. This way the system might be more user-friendly, trusted, and interactive between physicians and machine.

The accuracy of medical diagnosis could be seen from the perspectives of unfailing and consistent diagnosis based on clinical variables collected by pathologists which are presented by the sick patients. Given that manual diagnostic tools, give pathologists the freedom to exercise their domain expertise in decision making, the idea of using deep learning in clinical decision making becomes a threat to the physicians, due to its opaque (black box) analytics mechanism. This has necessitated the automation of clinical diagnostic workflow using deep learning which is eXplainable. The main focus of this research is to use some eXplainable AI techniques to give reasons for breast cancer diagnosis, from sampled histopathological images. This study, thus, develops an Explainable breast cancer prediction system for clinical transparency purposes. To achieve this, the study: (i) designed a model based on eXplainable AI (ii) implemented the model (iii) evaluated the performance of the proposed model with some sample data. and (iv) visualized and interpreted the model decisions using Explainable AI techniques.

## 2. Related Work

Various researchers all over the globe have begun to apply neural network methods to medical image analysis tasks, and are obtaining promising results. Several studies have been carried out using Deep Learning methods to Breast Cancer Diagnosis in histopathological images. New technology, based on artificial intelligence technology and some studies show that deep learning can be used for classification of breast cancer histopathology images [9]. Recent advancements in computing resources such as Graphical Processing Units (GPUs) and Tensor Processing Units (TPUs) and availability of large datasets, have made it possible to train larger and more complex neural networks. This has resulted in the design of several deep Convolutional Neural Networks (CNNs) architectures that are capable of accomplishing complex visual recognition tasks [10]. This shows that medicine, most especially medical image analysis, can benefit from deep learning technology through convolutional neural networks (CNNs). Here, we give a summary of some researches carried out on breast cancer diagnosis and explainability methods.

[11] proposed a Neural Network Based Algorithms for Diagnosis and Classification of Breast Cancer Tumor. The goal was to detect and classify cancer using an artificial Neural Network. The methodology employed was the Multi-Layer perceptron (MLP), using back propagation. Results showed that each layer of neurons in each network is linked to all previous neuron layers. The algorithm gave a performance accuracy of 82%. The method, however, proved to be computational intense, requires huge amount of data and lacks model explainability.

[12] classified breast cancer histology images using incremental boosting convolution networks. The goal was to detect breast cancer on breakHis dataset using CNN and boosting tree classifier. Results revealed that the methods are highly efficient for feature extraction and learn all level discriminant features of an input medical image, but poor in Bio imaging datasets when used on single Image Classifier. The model is classified as typically a black box model.

[13] classified Histopathological biopsy image using ensemble of deep learning Networks. The aim was to design a (Computer Aided Design) CAD for automatic binary classification of breast histology image. Multi model ensemble method for a multilayer perceptron classifier was used for the design. Results show that the proposed method was highly accurate and does automatic feature extraction. However, the model result lack transparency and is not suitable for small sized dataset.

[14] developed a new approach to computer-aided diagnosis scheme of breast mass classification using deep learning technology. The aim was to classify breast mass using CAD scheme. The methodology employed was deep learning. Results gave an automatic feature extraction with high accuracy. The method, however, lacks model explainability and is heavily dependent on the Region of Interest (ROI)

[15] gave a discriminative ensemble of histological hashing & class-Specific Manifold Learning for Multi-class Breast Carcinoma Taxonomy. The goal was to propose a model for both binary and multi-class breast cancer detection on BreakHis dataset. The method employed was an ensemble of histological hashing and class-specific manifold learning. Results showed high Accuracy, speed in detection and suitability for all both large and small sized data. However, the method lacks model transparency.

[16] study was on breast cancer multi-classification from histopathological images with structured deep learning model. The aim was to give a breast cancer multi-classification and patient level classification from histopathological images. A deep learning method based on GoogLeNet architecture was used for the image classification task, and a majority voting method was used for patient-level classification. Although, high classification accuracy was revealed, the results of the model was not visualized.

[17] gave a diagnosis of breast cancer using a combination of genetic algorithm and artificial neural network in medical infrared thermal imaging. The method employed was genetic algorithm and back-propagation neural network. Results revealed that the method can be used for patients with and without symptoms. However, the method lacks model explainability and needs improvement.

[18] gave a breast cancer histopathological image classification using convolutional neural networks. The goal was to classify breast histopathological images as benign or malignant using Convolutional Neural Network (CNN). The method proved to be very efficient, highly accurate with great success in detecting Region of Interest (ROI). The method, however, has difficulty in detecting early and 3<sup>rd</sup> stages of tumor. It is a blackbox model as it lacks explainability of classification.

[19] gave a classification of Histopathological Breast Cancer Images. The aim was to give a Histopathological Breast Cancer Image Classification by DNN Technique, guided by local clustering. The CNN method was employed. Although the method allowed improvement in the classification accuracy of the network, it is however slow in feature extraction and lacks model explainability.

While the reviewed literatures in above shows specific methods of detection, some were not evaluated on a large scale dataset, though they ended up producing viable clinical model

performances such as diagnostic accuracy, positive predictive value (PPV) etc. Existing approaches need to validate their robustness on a larger dataset. Without transparent interpretations on how they reached their diagnostic predictions, adopting the diagnostic model for a real clinical use case remains questionable by physicians and other medical experts. This simply means that the trust needed to carry out clinical actions based on the predictions from the model only exist if there is an explanation to the decision made by the model. Hence this study tends to cover that gap.

### 3. Methodology

The methodology used in this study is “Deep Learning”.

The following steps were followed in achieving the aim of this study, which is to visually interpret a histopathological breast cancer prediction results, for clinical transparency purposes.

- 1<sup>st</sup> Acquire a curated BreakHis dataset
- 2<sup>nd</sup> Divide data into 3 sets  
Training dataset (5,005)  
Test data (791)  
Validation dataset (2113)
- 3<sup>th</sup> Apply data augmentation using Pytorch vision transformation libraries,
- 4<sup>th</sup> Train a Resnet50 pretrained model using the concept of transfer learning.
- 5<sup>th</sup> Validate model accuracy with some sample data.
- 6<sup>th</sup> Visualize and interpret model decisions using Integrated gradient, GradientShap and Occlusion analysis from Pytorch Captum.

The following objectives are used to achieve this aim:

- ❖ Carry out Data Augmentation
- ❖ Train a Resnet pretrained model with BreakHis images.
- ❖ Test model accuracy with some sample data.
- ❖ Visualize and interpret model decisions using Integrated gradient & GradientShap & Occlusion.

### 4 Systems Analysis and Design

This section analyses the existing system and gives the model architecture for the prediction of two cancer labels (Malignant and Benign). The section also explains the dataset used, the concept of data augmentation, the training process, and the various interpretability algorithms used in explaining the model predictions.

#### 4.1 The Existing System

The existing systems, based on literature, are deep learning breast cancer prediction systems which are not explainable. They only have the capacity to predict if the breast tumor is malignant or benign without giving reasons for the prediction. These systems are plagued with the following drawbacks (i) understandability of the model prediction is limited to the developers alone (ii) accountability of the model cannot be measured if predictive errors occur. (iii) there is model mistrust as it lacks explainability to pathologists. (iv) not scalable, that is, the model performance cannot easily be improved on since its predictive performance cannot be accessed.

#### 4.2 The Proposed Model.

The proposed model is an Explainable breast cancer prediction system which was trained on ResNet50. It has the capacity to use Explainable AI (XAI) methods such as Integrated Gradient, GradientShap and Occlusion to visually interpret a cancer class. These explainable methods will enable pathologists to visually explain the results of breast cancer detection model. These methods tend to look out for image shapes, color or localization to draw its conclusion, depending on the researcher’s intentions.

##### 4.2.1 ResNet50 Pretrained Model Architecture.

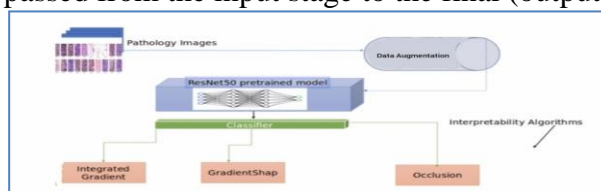
ResNet50 is a residual network that has 50 layers. It is a pretrained deep Convolutional Neural Network (CNN) for image recognition and supports residual learning; that is, it extracts features learnt from inputs of a layer using shortcut connections [20]. The idea behind the ResNet50 model was to develop a deep learning model with less training error, by adding shortcut connections also known as the 'residual connections' to the network to avoid loss of data during training and to boost the model performance [21]. The model has more than 20 million parameters for training which makes it suitable to build a deeper network. It is also known to have excellent performance in image recognition with lesser error rates compared to other pretrained models like VGG16, EfficientNet-BO, etc. It achieved a 3.57% error rate on the ImageNet test set [21]. The major advantage of residual connections in ResNet architecture is that during training, the knowledge acquired by the connections is preserved and it also speeds up the training time of the model by increasing the capacity of the network. Figure 1 shows a ResNet50 pretrained model adopted for this research.



**Figure 1: Resnet50 pretrained Model Architecture [21] (He et al., 2016).**

#### 4.2.2 Proposed Model Architecture.

The ResNet50 model is adopted for this research. ResNet50 Model is pretrained on a portion of ImageNet database and can group images into 1000 object categories. It is best for image recognition hence its usage in this research. Figure 2 shows the stages of the proposed model Architecture as the pathology images are passed from the input stage to the final (output) stage.



**Figure 2: Proposed Model Architecture**

#### 4.3 Benefits of XAI

The benefits of XAI include the following : (i) it helps in validating models predictions and also for gaining new insights into any new task. (ii) XAI helps in building trust by strengthening the stability, predictability of interpretable models (iii) it helps in improving model performance by understanding how models function (iv) it creates room for identification and correction of errors (v) it helps in retaining control over AI performance

#### 4.4 Perspectives of Explainability

Model explainability can be seen from two perspectives [22]:-

- i) Non-attribution based methods
- ii) Attribution based methods

**Non attribution based method:** - Non attribution based methods is domain specific. It deals with explainability problems by creating a method and authenticating it on a given problem rather than performing a different analysis using previously existing attributions based methods. Examples include attention based techniques, concept vector, expert knowledge etc. It is usually used on a post-hoc model (i.e. a model that has been built already).

**Attribution based method:** - It is a method for explaining the output of a DNN (deep neural network) model by considering the dominant features in the model that led to its prediction. In a classification problem the attribution method aims at adding features to the input to make the output neuron of the correct class. The features with a positive contribution to the activation of the target

neuron are marked in a specific color while those negatively affecting the activation are marked in a different color depending on the training mechanism. The commonly used attribution methods are perturbation based methods (such as occlusion), back propagation method, gradient based methods (such as integrated gradient, gradientShap and saliency map e.t.c) and surrogate method. In this study, three attribution based methods are used in explaining breast cancer diagnosis which includes: the gradient based methods (Integrated Gradient and GradientShaps) and the perturbation based method (Occlusion). These algorithms visualize breast cancer diagnostic model by using derivatives from the input class to predict the output class (gradient based method) and also hiding the irrelevant features in the image in order to predict the output class (perturbation based method).

The format of explanation according to [23] is as follows:-

1. **Explanation by Analysis of natural language statements**:-It describes the elements and context that make up a language statement, e.g. Text (Statements, Narratives or Stories, Answers to queries, Human machine dialogs).
2. **Explanation by Visualizations**:- This directly highlights the portions of the raw pixels that support a choice and allow viewers to form their own perceptual understanding. Example is heat maps.
3. **Explanation based on Cases**:- This involves specific examples or stories that support the choices made.
4. **Explanation based on Rejections of alternative choices**: - This involves arguing against less preferred answers based on analytics, cases, and data.

Explanation by Visualization using color variation and positioning of image features is our focus in this research, where a portion of pixel is highlighted as either Benign or Malignant.

#### 4.5 Data Acquisition.

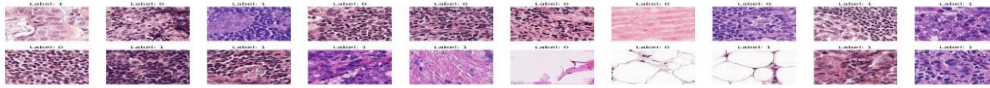
Data acquisition involves capturing input data for a model to work with. In this study, Breast Cancer Histopathological Database (BreakHis) was used. It comprises of pathological images, which are resident in kaggle repository. it contains 7,909 images of breast tumor tissue which includes 2,480 benign and 5,429 malignant sample (700 X 460 pixels, 3-channel RGB, 8-bit depth in each channel, PNG format),with the following magnifying factors (40X, 100X, 200X, and 400X). These sample images were collected from 82 patients, including those with clinical indication of breast cancer that were referred to P&D Laboratory in Parana, Brazil between January to December 2014. The recent BreakHis version, was collected by Surgical open biopsy (SOB) method, which cleans up the datasets by removing larger portion of irrelevant tissues making it curated. These images were labeled by pathologists of the P&D laboratory and ensured that no two images have same structure and label. Table 3 gives a breakdown of the BreakHis data distribution.

**Table 3: Data distribution of the breakHis dataset, by class and magnification.**

Magnification	40x	100x	200x	400x	Total	Patients sampled
Benign	625	644	623	588	2,480	24
Malignant	1,370	1,437	1390	1,232	5,429	58
<b>Total</b>	<b>1,995</b>	<b>2,081</b>	<b>2,013</b>	<b>1,820</b>	<b>7,909</b>	<b>82</b>

#### 4.6 Dataset Splitting

Data splitting involves sharing available data into portions for processing. The data from Table 3 were sampled as follows. Training dataset (5,005), Test data (791) and Validation data (2113). However, before the algorithm was built to model the data, random visualization was carried out on 20 samples taken from the BreakHis dataset (displayed in figure 3). Each pathological image was labeled either as 0 or 1. Images labeled 1 are malignant while those labeled 0 are benign.



**Figure. 3.** Randomly visualized images of the Pathology images.

Exploratory analysis was performed on these data, just to understand the pattern variation and the quality of the data labeling.

#### 4.7 Data Augmentation.

**Data Augmentation:-** It is the process of generating more data from limited data using different orientation in order to create new data and avoid overfitting. It is commonly used in convolutional neural networks (CNN) to improve model result. It is achieved using Pytorch vision transformation libraries which have several list of arguments used for training data.

Data augmentation can be classified under regularization method, which is used in adjusting the learning algorithm for the model to generalize effectively; this improves the model performance on the test data. What it actually does is to increase the actual size of the training dataset, by producing a transformed copy of each image, resulting in times 2 or 3 of dataset size. The following list describes the used data augmentation in this study.

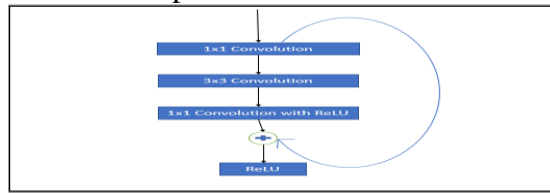
- ❖ **Image Resizing:-** Images were resized to 224 x 224 pixels. This does not affect the quality of the image, but rather reduces the height or width of the image.
- ❖ **Image rotation:-** Images were randomly rotated by 90 degrees several times. In our implementation, we set the probability to 0.5. This means there is a 50% chance that a given image will be rotated 90 degrees.
- ❖ **Image Transposition:-** Rows and columns of the image matrix were swapped. Probability of applying this on a given image was set to 0.5.
- ❖ **Images flipping:-** Flip the input image either horizontally, vertically or both horizontally and vertically, with a probability set to 0.5.
- ❖ **Hue Saturation Value:-** Randomly change color appearance of the input image, this also regulates the saturation and value. This result in a color shift of the image and was set to 0.5 probability of an image being picked.
- ❖ **Gaussian Noise:-** It is a statistical noise having a probability density function (PDF) equal to that of the normal distribution, which is also known as the Gaussian distribution, this is applied to the input image, with a probability of 0.5.
- ❖ **Random Affine Transform:-** Translate, scale and rotate the input image with a probability of 0.5. The translation is within the x- and y-axis of the image.
- ❖ **Brightness and Contrast:-** Randomly change brightness and contrast of the input image with a probability of 0.5
- ❖ **Blur:-** Blur the input image using a random-sized kernel. The probability was set to 0.5.
- ❖ **Sharpen:-** Sharpen the input image and overlay the results with the original image. The probability of sharpening a given image was set to 0.5.

#### 4.8 Training A Resnet50 Pretrained Model

This section covers specific steps for the training of a Resnet50 model which will assist in the prediction of the images to be either benign or malignant. A ResNet50 pretrained model was adopted for the purpose of predicting a cancer class using the concept of transfer learning.

ResNet50 is a deep Convolutional Neural Network developed by [21]. It contains 50 layers, which supports residual learning by extracting features learnt from an input layer using shortcut connections known as skip connections [20]. These skip connections preserves the knowledge learnt during training and speeds up the training time to avoid loss of data. It has the advantage of higher

accuracy with histopathological images compared to other image recognition models. Figure 4 is an example of a residual block that does skip connection.



**Figure 4** An example of a residual block in ResNet50 [21].

Transfer learning is the allocation of task(s) from a pre-trained model to a new model [24]. In this research, a deep convolutional Neural Network model was adopted (i.e. ResNet 50 model) using Transfer learning technique to study breast cancer diagnosis in histopathological images, trained on the curated BreakHis dataset

#### 4.8.1 Parameters for Training the Model:

Table 3 represents the list of the parameters for training the model.

**Table 3: Parameters for training the model**

	Hyperparameter	Value Used	
<b>Model</b>	Input Size	224 x 224	
	Epochs	20	
	Batch Size	128	
	Learning Rate	1e-4(0.0001)	
	Gradient	Adam	
	Optimizer		
	Batch normalizer		
	Global Average pooling		
<b>Data Augmentation</b>	Horizontal Flip Probability	0.5 probability	
	Vertical Flip Probability	0.5 probability	
	Rotation Range	[-180, 180] degree	
	Gaussian Blur	0.5 probability	
	Random Grayscale		
<b>Regularization</b>	L2 regularization	0.01	
	Exponential learning rate	gamma = 0.9	
	Verbose	1	

#### 4.8.2 Training and Validation Steps

The step followed by [25] is adopted here for both training and validation.

##### ➤ Training Pseudocode

Training Steps

1. Set ResNet50 model to training mode



2. Initialize loss as a list
3. Initialize variable for correct prediction
4. Iterate over batches in the data loader for training
5. Get the input images and their labels from the dataset
6. Move the input images and their labels to GPU device
7. Zero the parameter gradient
8. Execute a feed forward training pass over the dataset
9. Hold the cross entropy loss function in a variable "loss"
10. Hold the correct predictions in a variable in step 3, by taking the sum of all predicted labels and compare with the actual image label
11. Add the loss to the list of accumulated loss
12. Execute a backpropagation in order to train backward by trying to reduce the loss
13. Update the parameters
14. Calculate the accuracy by dividing the number of correct predictions by the total number of predictions
15. Calculate the average loss over all batches
16. Return the average loss and the accuracy as output

➤ **Validation Pseudocode**

1. Set ResNet50 to evaluation mode
2. Initialize loss as a list
3. Set model to inference mode
4. Iterate over batches in the data loader for evaluation
5. Get the input images and their labels from the dataset
6. Move input images and their labels to GPU device
7. Execute a feed forward training pass over the dataset
8. Hold the cross entropy loss function in a variable "loss"
9. Take the sum of predicted labels and compare with the actual labels to get correct predictions
10. Add the loss to the list of accumulated losses
11. Calculate validation accuracy by dividing the number of correct predictions by the total number of predictions
12. Calculate the average loss over all batches
13. Return the average loss and the validation accuracy

#### **4.9 eXplainability Algorithms.**

eXplainability in machine learning is the extent to which a model can be understood easily by humans. The aim of eXplaining a model is simply to find out the reason(s) why a model made certain predictions and to increase trust of a model prediction. Model eXplainability is important because it gives the leverage for model to be predicted easily and errors detection [26]. To visualize the prediction capability of the proposed model, three interpretable algorithms were used, integrated gradient(IG), GradientShap and Occlusion. For easy implementation, Pytorch model interpretation framework called Captum was used.

#### **4.10 Integrated Gradient (IG)**

Integrated gradient is a technique for interpreting a Deep Neural Network (DNN) which visualizes its input feature importance that contributes to the model's prediction. It does this by calculating the gradient of the predicted output to the input features. It visualizes the prediction by inputting a neutral image into the network, it then add features gradually to the image to increase the image intensity, then the gradients are calculated to ascertain which pixel affects the prediction most [27]. This technique is used for justifying a NN by mapping its predictions to input neurons. IG ensures

no changes are made to the original DNN as it predicts the output to the input image features. It is used for model accuracy metrics, model debugging and feature extraction. Integrated gradient has many use cases including explaining feature importance, recognizing non uniform distribution in a dataset, and model performance debugging. It is computationally advantageous in that, it can accommodate images with larger input pixels. In this research two basic characteristics were considered to prove the reliability of the model predictions. They include Sensitivity and Implementation Invariance.

Generally, integrated gradient is calculated as follows:

$$IG_i(X) = (x_i - x'_i) \times \int_{a=0}^1 \frac{df(x'+a(x-x'))}{dx_i} d\alpha \quad [27]. \quad (1)$$

Where:

$i$  is single pixel

$x$  is the input image

$x'$  is the baseline image

$x_i$  input image along the  $i^{th}$  dimension

$x'_i$  baseline image along the  $i^{th}$  dimension

$a$  is the path from the baseline to the input value

$df/dx_i$  = gradient of model prediction

Equation 1 defines an integrated gradient along an  $i^{th}$  dimension for an input image  $x$  and a baseline image  $x'$ .

From equation 1, the function ( $f$ ) is a differentiable function which acts on the input image ( $x$ ) to produce an output( $X$ )

- **Sensitivity:** - It is an axiom which checks if the baseline image ( $x'$ ) and input image ( $x$ ) are different in the single variable, including having a different output. If these conditions are meant, then that variable should receive some attributions [27]. This simply means an insensitive variable or a variable that has no output changes, get no attribution. This is expressed in equation 2

$$\text{If } \chi \neq 0 \text{ and } F(\chi_i) \neq 0 \quad [27]$$

therefore the attribution to that feature will be non-zero.

Where:

$\chi$  is the input image

$\chi_i$  is the baseline image

$F(\chi_i)$  is the function of the baseline image.

- **Implementation Invariance:** With implementation invariance, two neural network models compute the same mathematical function, no matter their differences in implementation. The attribution of all features should always be the same and its output shouldn't depend on architecture of the Neural Network.

#### 4.11 GradientShap.

GradientSHAP is a linear model explanation algorithm that approximates SHAP value by calculating the expectations of gradients by sampling randomly from the distribution of baselines. It adds white noise to each input image a couple of times and randomly selects a baseline from the baseline distribution and selects a random point along the line between the baseline and its input; it then computes the gradient of the outputs with respect to those selected random points. The final SHAP values therefore will be the expected values of gradients multiplied by the inputs baselines. SHAP (SHapley Additive exPlanations) is developed by [28]. It helps to interpret the prediction of

an instance  $x$ , which could be image feature, by computing the contribution of each feature to the prediction.

In some sense also, GradientShap could be viewed as an approximation of Integrated Gradient by computing the expectations of gradients for different baselines, since it's also a gradient based algorithm.

Therefore to interpret an input image such as histopathology image, pixels can be grouped to super pixels and the prediction distributed among the images.

One of the advantages of SHAP is that the Shapley value interpretation is represented as an additive feature attribution method, which is a linear model.

**Additive feature attribution methods:-**This method has an interpretation model that is a linear function of binary variables as shown in equation 3 [29]

$$g(z') = \phi_0 + \sum_{i=0}^M \phi_i z'_i \quad [29] \dots \dots \dots (3)$$

Where

$g$  is model interpretation

$z'$  is coalition vector,

$\phi$  is feature attribution for the feature

$M$  is the maximum number of simplified input features or coalition size

$i$  is the Shapley value.

To compute Shapley values, we first imagine that only some features values are present (present features) and some are not (absent features).  $\phi$  is calculated with a value that represents the linear model of coalition. For  $y$ , which is an instance of interest, the simplified input features  $y'$  is a vector of all 1's, meaning all feature values are "present features". Equation 3 can be simplified as shown in equation 4

$$g(y') = \phi_0 + \sum_{i=1}^M \phi_i \quad [29] \dots \dots \dots (4)$$

Where  $g(y')$  is the model interpretation for the input image.

Since SHAP calculates shapley values it is considered to be very efficient, symmetric and additive. There are three advantageous properties of additive feature attribution methods that have close familiarity with the Shapley estimation methods; they include (i) Local accuracy (ii) Missingness (iii) Consistency

**(Property 1)**

**Local accuracy:** This is the first desirable property. This property, demands the interpretation model to at least match the output of the function  $f$  for the coalition vector or simplified input image  $y'$  (which corresponds to the original input image  $y$ ), when approximating the original model of function,  $f$ , for a specific input image  $x$ . This is as illustrated in equation 5

$$f(y) = g(y') = \phi_0 + \sum_{i=0}^M \phi_i y'_i \quad [29] \dots \dots \dots (5)$$

The interpretation model  $g(y')$  matches the original model  $f(y)$  when  $y = h_y(y')$ , where  $\phi_0 = f(h_y(0))$  which represents the model output with all simplified inputs toggled off or disable.

Where

$f(y)$  is the original model

$g(y')$  is the interpretation model

$f(h_y(0))$  is the model output

**(Property 2)**

**Missingness:** This explains that any feature which contributes to a prediction should be given an attribution value of 1 and if a feature is missing it gets an attribution of 0 as shown in equation 6

$$y'_i = 0 \Rightarrow \phi_i = 0, \quad [29] \dots \dots \dots (6)$$

where  $y'_i$  denotes a simplified input image,

0 is absent of a feature value.

**(Property 3)**

**Consistency:** The consistency property states that if the contribution of a model input value increases or remains constant, the Shapley value also increases or remains constant as well irrespective of other features. This is as illustrated in equation 7

For instance.

$$\text{Let } f_y(z') = f(h_y(z')) \quad [29] \dots \dots \dots (7)$$

and  $z' / i$  indicates that  $z'_i = 0$ .

For any two models  $f$  and  $f'$  that satisfies

$$f'_y(z') - f'_y(z' \setminus i) \geq f_y(z') - f_y(z' \setminus i) \quad [29] \dots \dots \dots (8)$$

for all inputs  $z' \in \{0, 1\}^M$ , then

$$\phi_i(f', y) \geq \phi_i(f, y) \quad [29] \dots \dots \dots (9)$$

**4.12 Occlusion**

Occlusion method is most useful in cases such as image processing, where pixels in an adjacent rectangular region are likely to be highly related. Its sensitivity is a simple technique for understanding which parts of an input image are most important for a deep network's classification. In this research, we sought to know which pixels of the histopathology data is most important or which path plays a major role in the classification of benign and malignant features. A network's sensitivity to occlusion can be measured in different regions of the data using small perturbations of the data. A perturbation based occlusion approach to compute attribution, involves replacing each adjacent rectangular region with a given baseline image, and computing the difference in output [30]. For features located in multiple regions of the image, the corresponding output differences are averaged to compute the attribution for that feature. The simplicity of occlusion makes it very easy to implement.

**5 Implementation, Results and Discussion.**

This section discusses the results of the implementation of the proposed model. Model training and validation output logs, model accuracy, loss function graphs and the model prediction interpretations, which reveal what features are responsible for the classification of histopathological images.

**5.1 Implementation**

ResNet50 was trained on the 40x magnification version of the BreakHis dataset and the interpretability algorithms were tested on the 400x magnification version of the dataset. This helps in ascertaining how well the model generalizes its understanding of benign and malignant features in images, even at different microscopic levels.

**5.2 Model Training Settings**

Table 4. shows all the parameters and values that were used for training the model. After experimenting with several values, the best result is what is displayed here.

**Table 4: Parameters for Model Training**

Training Parameters	Value
Batch size	128

Number of epoch	20
Learning rate	1e-4
Loss function	Cross entropy
Dropout	0.5 probability

### 5.2.1 Training logs

Table 5 shows the output logs generated during the training of the model. From these logs, it could be seen that while the training and validation accuracy increased, their losses decreased over time.

**Table 5: Training Logs Of Model**

Epoch	Logs	Training(%)	Validation(%)
1/20	Accuracy	83.79	70.37
	Loss	39.33	61.12
2/20	Accuracy	85.76	71.21
	Loss	37.61	66.91
3/20	Accuracy	87.74	87.10
	Loss	29.05	66.95
4/20	Accuracy	92.71	85.01
	Loss	21.99	50.09
5/20	Accuracy	93.00	83.12
	Loss	19.96	41.22
6/20	Accuracy	94.44	91.20
	Loss	17.21	39.23
7/20	Accuracy	95.25	75.29
	Loss	17.20	91.25
8/20	Accuracy	96.05	93.14
	Loss	18.16	29.99
9/20	Accuracy	96.07	90.46
	Loss	17.04	30.11
10/20	Accuracy	97.08	94.10
	Loss	15.12	11.21
11/20	Accuracy	91.34	87.94
	Loss	14.93	39.13
12/20	Accuracy	96.39	75.99
	Loss	13.31	59.36
13/20	Accuracy	96.41	78.32
	Loss	11.94	31.10
14/20	Accuracy	96.54	81.73
	Loss	11.54	29.91
15/20	Accuracy	96.75	96.00
	Loss	10.38	19.21
16/20	Accuracy	96.91	76.94
	Loss	10.70	82.14
17/20	Accuracy	97.33	76.97
	Loss	10.36	90.12
18/20	Accuracy	97.94	95.33
	Loss	10.32	89.52

19/20	Accuracy Loss	98.95 10.27	96.75 10.39
20/20	Accuracy Loss	98.98 10.22	<b>96.84</b> 10.39

**Best validation accuracy: 96.84%**

Figure 5 depicts the model accuracy and loss function figures for Resnet50 after training

```

#Loading the test data using Image Data Generator
test_gen = datagen.flow_from_directory("../Cancer_test/", target_size=(128,128), class_mode="binary", batch_size=1, shuffle=False)

Found 791 images belonging to 2 classes.

+ Code + Markdown

pred = model.evaluate(test_gen)

791/791 [=====] - 21s 26ms/step - loss: 0.1039 - accuracy: 0.9684
    
```

Figure 5: Model Accuracy and loss function figures for Resnet50 after training

**5.3 Model Accuracy and Loss Function Graph Plot.**

Model accuracy is a measure of the overall effectiveness of the model. [31] defined accuracy as the percentage of correct predictions for the test data, and calculated as:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{all Predictions}} \quad [31] \tag{10}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad [32] \tag{11}$$

Where *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives, and *FN* = False Negatives.

The following equations, given in 11 to 14 can be used to calculate model accuracy:

$$\text{Accurate predictions} = \text{True predictions} = \text{True Positive} + \text{True Negative} \quad . \tag{12}$$

$$\text{Inaccurate predictions} = \text{wrong predictions} = \text{False Positive} + \text{False Negative} \quad . \tag{13}$$

$$\text{All prediction} = \text{True predictions} + \text{wrong prediction} \quad . \tag{14}$$

$$\text{Accuracy} = \frac{\text{TruePredictions}}{\text{TruePredictions} + \text{wrongPredictions}} * 100 \% \quad . \tag{15}$$

Therefore our best validation accuracy is  $0.9684 \times 100 = 96.84\%$  Accuracy

For a clearer understanding of the model output logs, a graph showing the Model Accuracy and loss function graph was plotted as shown in figure 5, where the blue line represents the training curve and the orange line represent the validation curve.

Figure 6 shows the model validation accuracy and the loss function value obtained during training. The validation loss of ResNet50 is 10.39%, the training accuracy is 98.98% and the validation accuracy is 96.84%. Sometimes the reason why Training Accuracy is higher than that of Validation Accuracy is because a larger percentage of the entire dataset was used for training. It was also observed, that while the epoch increases, the training and validation accuracies increases as well while the loss reduces.

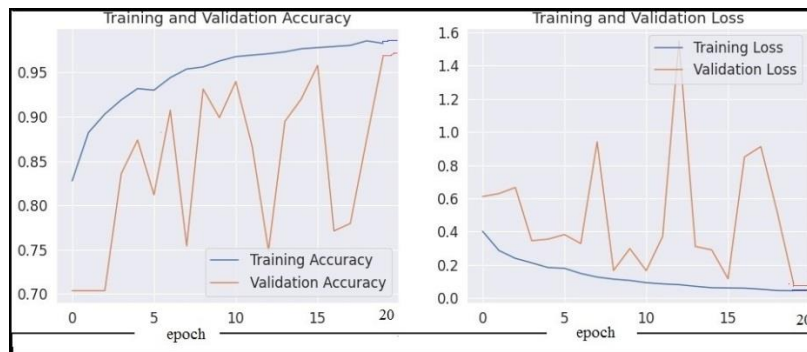


Figure 6: Model Accuracy and loss function graph for Resnet50

#### 5.4 Model Prediction and Interpretability of Results

Model prediction and interpretability of results look into the visual results from three interpretability algorithms, Integrated Gradient (IG), GradientShap (GS) and Occlusion, which are meant to visually interpret histopathology images and also give the ability to see how these algorithms uniquely explain the predictions. Integrated Gradient based its predictions by mapping similar features which are of utmost importance to the input image, analyzing it and drawing a conclusion on it. GradientShap interpret the prediction of an input image by computing the contribution of each pixel to the prediction. It extracts Shapley values from coalitional game theory and allocates properly the payoff among the features visualized. Occlusion on the other hand, removes the irrelevant features of the images, making visualization easy. A comparison will be made between the interpretability algorithms based of their time of execution and visual clarity regardless of the model accuracy.

The images for visualization were conducted on 400x magnified breakHis dataset.

##### 5.4.1 Interpretability Testing Techniques

Interpretability testing techniques are techniques used to explain a deep learning model, taking into consideration its shapes, location or color variation.

In this research, two expert labels were randomly picked with 400x magnified pathological images that were not part of the training data, to see which region of the image is regarded by the model to be either benign or malignant. From the images selected, the model predicted the 2 images correctly (true positive). Since the diagnosis was already known before hand, it wasn't difficult to detect if the model was predicting correctly or not by showing the regions using Integrated Gradient, GradientShap and Occlusion.

#### 5.4.1.1 Analyzing the predictions.

Figure 7(a) shows an image of BreakHis dataset that has been diagnosed by an expert, while figure 7(b) shows that the predicted image is benign due to the concentration of black spots as the scale moves from 0.0 to 1.0

#### DESCRIPTIONS

**Expert label:** -Benign; **Magnification:** - 400x; **Model prediction:** - Benign, with 89% model accuracy; this is as a result of color concentration on the region of interest.

#### **Algorithm 1: Integrated Gradient.**

**Explanation:** Using Integrated Gradient (IG), we were able to attribute the model prediction of benign to the input image in form of black spots. That is, regions where the model concentrated the most while analyzing the image, are attributed black spots looking as plain as possible from the input image. These regions are where the model sees benign. The more concentrated these black spots are in the visualized image, the more the likelihood that the image is benign. It was not difficult to attribute the model prediction of benign to the input image as the algorithm extracted rules from the network, debug the performance of the model and identified the important features where it is concentrated and based its predictions on it. The attributed black spots from observation may be as a result of lumps concentrated on a particular region of the image in reality. These regions are where the model sees benign. Figure 7b shows that the predicted image is benign due to the concentration of black spots as the scale moves from 0.0 to 1.0. The more concentrated these black spots are in any part of the image, the more the likelihood that the image is benign in that part. From the visualization depicted in (Figure 7b), the bottom right-hand side of the input image is where the model picked as the benign signal, with approximately 63.1 seconds of execution time.

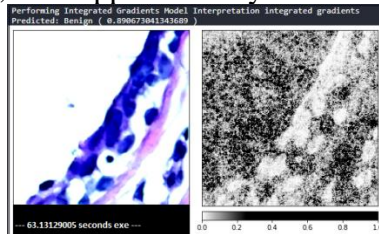


Figure 7a      Figure 7b

Figure: 7 (a): Expert Labeled Benign Image of BreakHis dataset

Figure: 7 (b): Integrated Gradient labeled Benign Image of BreakHis dataset.

#### **Algorithm 2: GradientShap**

**Explanation:** In applying GradientShap (GS), Shapley values are to be calculated and this gives the contribution of each feature to the model prediction. These Shapley values assign importance to each feature that makes up the model prediction of being benign. These features are what the model sees while recognizing benign. Features with high importance are represented with black spots. Just like Integrated gradient, the more concentrated these black are in a particular region that is closely fitted, the higher the chances that the image is benign. Figure: 8 (a) shows Expert Labeled Benign Image of BreakHis dataset, while figure 8(b) shows features of high importance which is the concentrated black spots seen at the bottom right corner of the visualized image with an execution time of approximately 62.9 seconds.

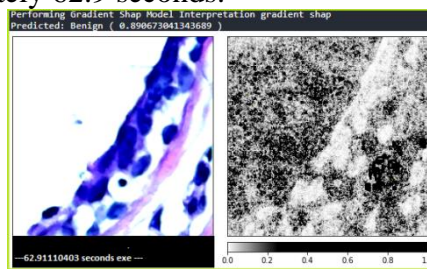




Figure 8 a      Figure 8b

Figure: 8 (a): Expert Labeled Benign Image of BreakHis dataset

Figure: 8 (b): GradientShap labeled Benign Image of BreakHis dataset.

### **Algorithm 3 : Occlusion**

**Explanation:** Occlusion allows the estimation of the region of the image that is critical for the model's decision, in order to ascertain the area that is benign or malignant. This it does by hiding the irrelevant parts of the image and measuring how the decision changes. This is achieved by running a 15x15 pixel sliding window across the image at each point, with a baseline of 0. For features located in multiple regions of the image, the corresponding output differences are averaged to compute the attribution for that feature. Figure 9(b) illustrates this; the visualized areas with thicker green square concentrations indicate high likelihood of the presence of benign as the scale moves from 0.0 to 1.0. Performing Occlusion interpretability algorithm, the model predicted that the image is Benign with an execution time of 59.44seconds. Figure: 9(a), however, shows Expert Labeled Benign Image of BreakHis dataset

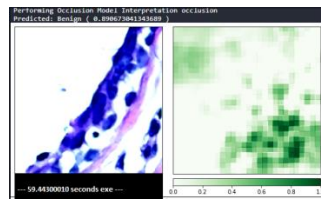


Figure 9a      Figure 9b

Figure: 9(a): Expert Labeled Benign Image of BreakHis dataset

Figure: 9 (b): Occlusion labeled Benign Image of BreakHis dataset

### **Image two Result**

Figure: 10 shows an expert labeled malignant image of BreakHis dataset.

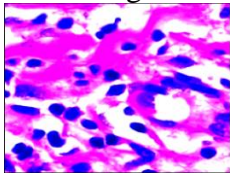


Figure: 10: Expert labeled Malignant Image of BreakHis dataset

**Expert label:- Malignant; Magnification:- 400x; Model prediction:** Model predicted that the image is malignant with 95% accuracy based on its coloration at the region of interest and how the pixels were sparsely dispersed on different locations of the image.

### **Algorithm1: Integrated Gradient (IG)**

**Explanation:** Integrated gradient calculates the essential areas of a model output for the predicted class with respect to the input image pixels. Integrated Gradient (IG) attribute the model prediction of malignancy to the input image in the form of concentrated black spots that are sparsely scattered. The regions where the model concentrated the most, while analyzing the image, are attributed black spots and these spots are also present in other locations of the image. These regions are where the model sees malignancy. Figure 11 (a) shows Expert Labeled Malignant Image of BreakHis Dataset, while Figure 11b shows that the predicted image is malignant due to the concentration of black spots as the scale moves from 0.0 to 1.0. The more concentrated these black spots are in different parts of the image, the more likely that particular image is malignant. From the visualization depicted in (figure 11b), the middle left part of the image is where the model picked as the malignant signals, with an execution time of 62.1seconds.

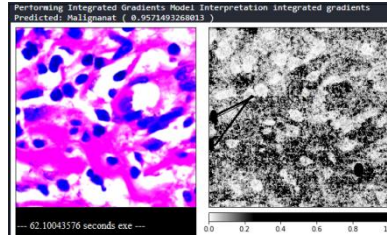


Figure 11a      Figure 11b

Figure: 11 (a): Expert Labeled Malignant Image of BreakHis Dataset

Figure: 11 (b): Integrated gradient labeled Malignant Image of BreakHis dataset

**Algorithm 2: GradientShap (GS)**

**Explanation** Gradient Shap (GS) calculates shapley values which assigns importance to each feature that make up the model prediction. These features are what our model sees while recognizing malignancy. Features with high importance are represented with black spots. Just like Integrated gradient, the more concentrated these black spots are, the higher the chances that the image is malignant. Figure: 12 (a) shows Expert Labeled Malignant Image of BreakHis Dataset, while figure 12 (b) shows features of high importance which is the concentrated black spots seen thus predicting the image as malignant with an execution time of 61.90 seconds. GradientShap has more noise compared to other algorithms, so visualization is usually not clear.

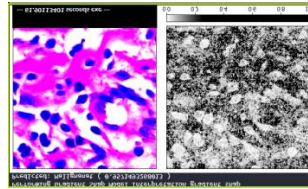


Figure 12a

Figure 12 b

Figure: 12 (a): Expert Labeled Malignant Image of BreakHis Dataset

Figure: 12 (b): GradientShap labeled Malignant Image of BreakHis dataset

**Algorithm 3: Occlusion**

**Explanation:** Occlusion is noiseless, that is, it hides the irrelevant parts of the image during interpretation and measures how the decision changes. It allows the estimation of the region of the image that is critical for the model's prediction. This is achieved by running a 15x15 pixel sliding window across the image, and at each point. For features located in multiple regions of the image, the corresponding output differences are averaged to compute the attribution for that feature. Figure 13 (b) illustrates this; the visualization areas with thicker green square concentrations indicate high presence of malignancy, as the scale moves from 0.0 to 1.0, with an execution time of 60.1 seconds. Figure: 13 (a), however, shows Expert Labeled Malignant Image of BreakHis dataset.

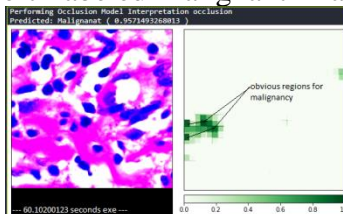


Figure 13 a      Figure 13b

Figure: 13 (a): Expert Labeled Malignant Image of BreakHis dataset

Figure: 13 (b): Occlusion labeled Malignant Image of BreakHis dataset.

**5.5 Parameters Contributing To Model Accuracy**

Table 6 gives the Parameters contributing to Model's Accuracy

**Table 6: Parameters contributing to Model accuracy.**

Parameter	Value	Functions/Remark
Early stopping	-	To avoid overfitting.

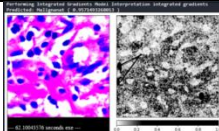
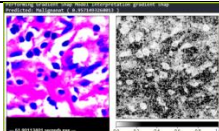
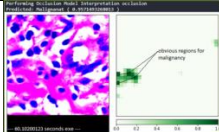
Regularizer (L2 Regularizer)	0.01	L2 regularizer sums the square of all feature weights. Discourages the complexity of deep learning model. Penalizes the loss function. Brings the value of weight closer to 0.
Dropout	0.5	To avoid overfitting.
Optimizer (Adaptive gradient decent)	Adam Optimizer	It updates the learning rate of every individual parameter during gradient update.
Batch Normalizer	128	Helps in model training speed and stabilized distribution of input data during training.

## 5.6 Discussions

### 5.6.1 Comparison of Applied Algorithms.

Table 7 gives the comparison of the interpretable algorithms applied on the model. From the analysis, Occlusion is seen to be the best, putting execution time and visual clarity into consideration. Occlusion visualizes the image better and faster than Integrated Gradient and GradientShap with less noise (i.e. unnecessary details) and faster execution time.

**Table 7: Comparison of applied Algorithms.**

Algorithm	Image	Time of execution	Analysis
Integrated Gradient(IG)		62.1 sec.	<ul style="list-style-type: none"> <li>- Model Predicted well but at a glance the point of interest cannot be easily visualized.</li> <li>- Too much noise in visualized Image.</li> <li>- More time of execution.</li> </ul>
Gradientshap(GS)		61.90 sec.	<ul style="list-style-type: none"> <li>- It has good model prediction.</li> <li>- More noise compared to other eXplainable algorithms</li> <li>- Time of execution better than Integrated Gradient.</li> </ul>
Occlusion		60.1 sec	<ul style="list-style-type: none"> <li>- Image well visualized with less noise.</li> <li>- Faster execution time.</li> </ul>

### 5.6.2 Comparison with other related Research work

Results from some selected existing work that used the same method were compared with the results of this study and the proposed model was found to have the highest accuracy and it visually explained the results of the model unlike other literatures. Table 8 gives the comparison of the proposed model with other related work

**Table 8: Comparing developed model with other Existing Literatures.**

AUTHOR(S)	AIM	MODEL	ACCURACY	MODEL RESULT EXPLAINED?	EXPLAINABLE ALGORITHM(S) USED
Developed model	To classify & visually interpret breast histopathological images.	CNN	96.84%	Yes	Integrated Gradient, GradientShap & Occlusion
[18].Spanhol., <i>et al</i> (2016).	Classification of breast histopathological images as benign or malignant.	CNN	80.4%	No	Nil
[19].Nahid <i>et al.</i> , (2018).	Classification of breast cancer images.	CNN	91%	No	Nil
[33] Pereira <i>et al.</i> , (2018).	To diagnose and visualize brain tumor from MRI.	CNN	Nil	Yes	GradCam & Guided Backpropagation (GBP)

[33] used GradCam to produce a rough localization map of the important regions in brain image. In GradCam, visualization becomes difficult because of the scattered heatmap. This problem is also common with integrated gradient, but the introduction of noise tunnels, can smooth out these uncertainties. With Occlusion being part of our interpretability algorithms, we have more model prediction explanations than what the other authors have. Occlusion gives a more accurate feature location of a visualized image.

## 6. Conclusion and Recommendation

In this research, ResNet50 was trained on the BreakHis dataset using a deep learning method (CNN) to classify breast tumors as either benign or malignant. Additionally, Explainable Artificial Intelligence (XAI) techniques, such as Integrated Gradient (IG), GradientShap (GS), and Occlusion, were used to visually explain the presence of cancer labels in histopathological images. These images were interpreted using XAI techniques based on color variation and location. The more concentrated the black or green coloration in an image, the more likely that image is malignant or benign. Comparing the three interpretable algorithms, it was found that Occlusion visualizes better than Integrated Gradient and GradientShap because it has less noise, making it less strenuous for pathologists to draw their conclusions. The experiment achieved a validation accuracy of 96.84%, outperforming related literature due to its high predictive value and visual results explanation. This research adds to the body of knowledge by making the result of breast cancer diagnosis transparent through explainable AI techniques.

Is it possible to have 100% accuracy? Is there any better algorithm than Occlusion? Can these algorithms work perfectly well in larger datasets? These areas can be explored in future work towards a near perfect model that can support pathologists in breast cancer diagnosis and other stakeholders.

## References

- [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] Jayant S. Vaidya and Vivek Patkar (2016). Fast Facts: Early Breast Cancer Retrieved June 26, 2022 from <https://books.google.com.ng/books?>
- [5] R. L. Siegel, K. D. Miller and A. Jemal (2017). Cancer statistics, 2017. CA: A Cancer Journal of Clinicians.. Volume 67, Issue 1 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>
- [6] 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>
- [7] 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>.
- [8] 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\\_](https://www.researchgate.net/publication/40907393_)
- [9] 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>
- [10] 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>
- [11] I-S. Jung, D. Thapa, and G-N. Wang (2005). "Neural Network Based Algorithms for Diagnosis and Classification of Breast Cancer Tumor" In 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](https://doi.org/10.1007/11596448_15) Retrieved April 22, 2022 from [https://page-one.springer.com/pdf/preview/10.1007/11596448\\_15](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] Y. Qiu, S. Yan, R. R. Gundreddy, Y. Wang, S. Cheng, H. Liu and B. Zheng (2017). "A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology" J Xray Sci Technol. 2017; 25(5): 751–763. doi: 10.3233/XST-16226. Retrieved July 13, 2022 from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5647205/>
- [15] S. Pratiher and S. Chatteraj (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
- [16] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li & S. Li (2017).. "Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model". Scientific Reports 7(1), 4172. doi: 10.1038/s41598-017-04075-z Retrieved August 10, 2022 from <https://www.researchgate.net/publication/317833617>
- [17] J. Haddadnia, M. Hashemian and K. Hassanpour (2012). Diagnosis of Breast Cancer using a Combination of Genetic Algorithm and Artificial Neural Network in Medical Infrared Thermal Imaging" Iranian Journal of Medical Physics 9(4):265-274 Retrieved March 10, 2022 from <https://www.researchgate.net/publication/>
- [18] F. A. Spanhol, L. S. Oliveira, C. Petitjean and L. Heutte (2016). "Breast cancer histopathological image classification using convolutional neural networks". International Joint Conference on Neural Networks (IJCNN 2016), Vancouver, Canada. pp 1-9 DOI:10.1109/IJCNN.2016.7727519. Retrieved May 10, 2022 from <https://www.researchgate.net/publication/304158394> [19] A.-Al. Nahid., M. A. Mehrabi, and Y. Kong (2018). Histopathological Breast Cancer Image Classification by Deep Neural Network Techniques Guided by

- Local Clustering. Hindawi BioMed Research International. Volume 2018 , pp 1-20. <https://doi.org/10.1155/2018/2362108>
- [19] V. Narayanan (2019). Image Classifier using Resnet50 Deep Learning model (Python Flask in Azure) . Retrieved 4th September, 2022 from [www.medium/@venkinarayanan/tutorial-image-classifier-using-resnet50-deep-learning-model-python](http://www.medium/@venkinarayanan/tutorial-image-classifier-using-resnet50-deep-learning-model-python)
- [20] K. He, X. Zhang, S. Ren and J. Sun (2016). Deep residual learning for image recognition . In proceedings of the IEEE Conference of Computer Vision and pattern Recognition page 770-778. <https://doi.org/10.1109/CVPR.2016.90>
- [21] A. C. Gusmao, A. H. C. Correia, G. De Bona and F. G. Cozman (2018). Interpreting Embedding Models of Knowledge Bases: A Pedagogical Approach. 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden pp79-86. Retrieved July 22, 2022 from <http://sites.poli.usp.br/p/fabio.cozman/Publications/Article/gusmao-correia-bona-cozman-whi2018F.pdf>.
- [22] D. Gunning (2017). Explainable Artificial Intelligence (XAI). Retrieved February 10, 2022 from <https://nsarchive.gwu.edu/sites/default/files/documents/5794867/>
- [23] S. J. Pan, and Q. Yang (2010). A Survey on Transfer Learning. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 22(10), 1345–1359. doi:10.1109/TKDE.2009.191. Retrieved June 10, 2022 from <http://www-edlab.cs.umass.edu/cs689/reading/transfer-learning.pdf>
- [24] W. Samek (2017). “Evaluating the visualization of what a deep neural network has learned,” IEEE Trans. Neural Networks Learn. Syst.28(11), 2660–2673.
- [25] C. Molnar (2020). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable Retrieved 5th September 2022 from <https://christophm.github.io/interpretable-ml-book/>
- [26] M. Sundararajan, A. Taly and Q. Yan (2017). Axiomatic Attribution for Deep Networks . Proceedings of the 34th International Conference on Machine Learning, PMLR 70:3319-3328, 2017. Retrieved July 09, 2022 from <http://proceedings.mlr.press/v70/sundararajan17a/sundararajan17a.pdf>
- [27] S. M. Lundberg and S-I. Lee (2017) . A Unified Approach to Interpreting Model
- [28] Predictions. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA., Retrieved September 10, 2022 from <https://arxiv.org/pdf/1705.07874.pdf>
- [29] S. M. Lundberg , B. Nair , M. S. Vavilala , M. Horibe , M. J. Eisses , T. Adams , D. E Liston , D. K-W. Low , S-F Newman , J. Kim and S-I. Lee (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018 , 2(10): 749–760., DOI: 10.1038/s41551-018-0304-0
- [30] M. D. Zeiler and R. Fergus (2014). Visualizing and Understanding Convolutional Networks. European Conference on Computer Vision. ECCV 2014: Computer Vision – ECCV 2014 pp 818–833. Retrieved October 10, 2022 from [https://link.springer.com/chapter/10.1007/978-3-319-10590-1\\_53](https://link.springer.com/chapter/10.1007/978-3-319-10590-1_53)
- [31] J. Jordan (2017), “Evaluating a machine learning model” Retrieved October 10, 2022 from [www.jeremyjordan.me/evaluating-a-machine-learning-model/](http://www.jeremyjordan.me/evaluating-a-machine-learning-model/)
- [32] Classification: Accuracy (2022) Classification: Accuracy - Machine Learning Google Developers. Retrieved July 05, 2022 from <https://developers.google.com/machine-learning/crash-course/classification/>
- [33] S. Pereira, R. Meier, V. Alves, M. Reyes and C. A. Silva (2018) Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment. Paper presented at iMIMIC - Workshop on Interpretability of Machine Intelligence in Medical Image Computing. <https://doi.org/10.48550/arXiv.1809.09468>. Retrieved July 03, 2022 from <https://arxiv.org/abs/1809.09468>.