

Article

Not peer-reviewed version

Explainable AI for Interpretation of Ovarian Tumor Classification Using Custom ResNet60

[Srirupa Guha](#)*, [Ashwini Kodipalli](#)*, [Steven L. Fernandes](#)*, Santosh Dasar*

Posted Date: 28 May 2024

doi: 10.20944/preprints202405.1865.v1

Keywords: ResNet60; Explainable AI methods; LIME, Saliency Map; Occlusion Analysis; GRAD-CAM; SHAP; SmoothGrad; Interpretable AI



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Article

Explainable AI for Interpretation of Ovarian Tumor Classification Using Custom ResNet60

Srirupa Guha ^{1,*}, Ashwini Kodipalli ^{2,*}, Steven L Fernandes ^{3,*} and Santosh Dasar ^{4,*}

¹ Department of Computer Science and Engineering, National Institute of Technology Durgapur, India

² Department of Artificial Intelligence and Data Science, Global Academy of Technology Bangalore, India;

³ Department of Computer Science, Design, Journalism, Creighton University, Omaha, NE, USA

⁴ Department of Radiology, SDM College of Medical Sciences Dharwad, India

* Correspondence: srirupaguha18@gmail.com (S.G.); dr.ashwini.k@gat.ac.in (A.K.); stevenfernandes@creighton.edu (S.L.F.); dr.santoshdasar@sdmuniversity.edu.in (S.D.)

Abstract: Deep learning architectures like ResNet and Inception have produced accurate predictions for classifying benign and malignant tumors in the healthcare domain. This enables healthcare institutions to make data-driven decisions and potentially enable early detection of malignancy by employing computer-vision-based deep learning algorithms. These Convolutional Neural Network algorithms, in addition to requiring huge amounts of data, can identify higher and lower-level features that are significant while classifying tumors into benign or malignant. However, these algorithms have limitations concerning explainability and identifying the exact features that are of importance and contributing to the final classification of tumors as benign or malignant. In this paper, we implement several explainable AI techniques, namely, LIME, Saliency Map, Occlusion Analysis, Grad-CAM, SHAP, and Smooth Grad to interpret the results of a custom ResNet60 classifier in the classification of ovarian tumors as benign or malignant. The ResNet60 model attained an accuracy of 97.50% for the test dataset.

Keywords: ResNet60; explainable AI methods; LIME; Saliency Map; occlusion analysis; GRAD-CAM; SHAP; SmoothGrad; Interpretable AI

1. Introduction

Ovarian cancer is one of the most common types of cancer in women all around the world [1]. Early detection remains one of the best possible ways of cure. Due to recent advancements in the Computer Vision domain, healthcare experts are leveraging CNN-based architectures to classify tumors into benign or malignant based on MRI or CT scanned images, thus enabling them to make early diagnosis of the disease. In addition to domain knowledge, experts can now rely on the capability of artificial intelligence to detect signs of cancer at an early stage which would potentially transform the diagnosis of various types of cancer, with ovarian cancer being the most prominent. [2].

Recent advancements in deep learning convolutional neural network-based architectures have proven to be efficient in identifying higher and lower-level features during classification of tumors into benign and malignant categories at earlier stages of detection, which is nearly impossible to predict by the human eye [3]. These architectures provide enhanced performance when compared to traditional Machine Learning classification algorithms like Support Vector Machines and Logistic Regression due to the consumption of huge amounts of training data and the complexity of the network architecture. However, unlike traditional Machine Learning algorithms like Logistic Regression and Decision Trees, the deep learning CNN architectures provide lesser explainability and interpretability of the results despite outperforming these algorithms manyfold in terms of classification accuracy [4]. Hence it is crucial to understand the way CNN algorithms work for making a classification and the importance of the features considered for the final classification result for the healthcare experts to make a data-driven diagnosis with more confidence and prescribe the most effective treatment [5].

In this paper, we have developed a custom ResNet60 model for the classification of ovarian tumors into benign and malignant. This model was compared with several state-of-the-art architectures and has outperformed the other architectures with a classification accuracy of 97.50% on the ovarian tumor test dataset. Therefore, the results of the ResNet60 classifier are carried forward for interpretability. To

explain the results of the ResNet60 classifier, we have implemented the below explainable AI methods: LIME, Saliency Map, Occlusion Analysis, Grad-CAM, SHAP, SmoothGrad.

LIME is a model-agnostic method that aims at explaining the important features behind the classification results by creating several artificial data samples and thereby predicting the class of each such artificial data point. The Saliency Map highlights the pixels that most affect the class scores. Occlusion analysis evaluates the importance of each input dimension by analyzing the model performance with that input dimension missing and its impact on the model performance. Grad-CAM weighs the gradient in the final CNN layer against the output of this layer to determine which features are the determining factors behind the class predicted. SHAP is a feature importance technique while Smooth Grad adds a small Gaussian noise and highlights the important pixels contributing to the final predicted class across all the training images. Using these explainable AI methods, we interpret and explain the results of the ResNet60 by highlighting the important features and pixels considered by the ResNet60 classifier while classifying ovarian tumors into benign and malignant. Furthermore, it helps us understand the key characteristics of benign ovarian tumors, which distinguish them from malignant tumors.

The main contributions of this paper are summarized below:

- A novel ResNet60 architecture is proposed for the classification of ovarian tumors as benign and malignant.
- The ResNet60 classifier performs the classification task with 97.5% accuracy on the test dataset, which is higher than other state-of-the-art architectures implemented - GoogLeNet (Inception-v1), Inception-v4, VGG16, VGG19, ResNet50, EfficientNetB0.
- The ResNet60 model results are then provided for interpretation to the explainable AI methods LIME, Saliency Map, Grad-CAM, Occlusion Analysis, SHAP and SmoothGrad.
- The above explainable AI methods highlighted the features and regions of interest that were given most importance by the ResNet60 model in obtaining the final classification output, such as the shape and neighbouring area of the tumor.

The organization of the paper is as follows: Section 2 presents a literature review of various explainable AI approaches like LIME, SHAP, Grad-CAM, SmoothGrad for image classification in general as well as in the medical imaging domain. Section 3 discusses the gap in existing literature for leveraging explainable AI methods used in classifying ovarian tumor images and the motivation behind this work. Section 4 describes the source and distribution of the ovarian tumor CT scan image dataset used for the classification. Section 5 discusses the proposed ResNet60 architecture and an overview of the explainable AI methods used for the interpretation of the results. Section 6 presents a comparison of the performance of ResNet60 with state-of-the-art architectures and shows that ResNet60 yields the highest test accuracy in the classification of tumors and hence the results are interpreted further using the explainable AI methods as mentioned above. Section 7 discusses the key takeaways from this research work and the scope of future work in the same domain.

2. Related Work

2.1. Existing Work Explainability of Classification Models in the Medical Imaging Domain

In the healthcare domain, the use of AI to classify images of tumor obtained from CT scan or MRI scan, as benign or malignant is increasingly becoming a useful tool for early diagnosis of cancer. Also, with the advent of COVID-19, we see research work on classification of chest X-Ray images for presence or absence of COVID-19 or pneumonia. Some of these use cases are illustrated by the following research papers. For example, Manali Gupta et al. [6] implemented and evaluated the performance of a scratch CNN method with VGG-16 for classification of brain tumor MRI images as cancerous or non-cancerous. Samir S. Yadav et al. [7] evaluated convolutional neural network based architectures for classification of pneumonia presence on chest X-Ray images dataset. Although deep neural networks have proved to be effective in the medical image classification task, with increasing

number of layers, the training process can slow down and become less effective due to the problems of vanishing and exploding gradient. This problem is solved by introducing residual networks, commonly known as ResNet, which enhances the training of deep neural networks. Jiazhi Liang et al. [8], in his research paper, illustrated the ResNet architecture for image classification using deep neural network models. Once the CNN models are trained and used to make predictions on medical images, the next task is to explain the classification output. Due to the black box nature of neural networks, often it is difficult to explain the nature of the output and why the model predicted so. Hence explainable AI methods are crucial at this point to interpret and explain the classification predictions obtained from the neural network. Following are some surveys and review papers on explainable AI methods, used for reference. A.S. et al. [9] discussed some of the commonly used explainable AI methods like LIME and SHAP and several versions of the same. Baehrens, D. et al. [10] uses local explanation vectors for explaining classification results. Xu, F. et al. [11] does a detailed survey of the history and various methods in the field of explainable AI. Yang, W. et al. [12] discussed various explainable AI approaches along with their limitations and use cases. Samek, W. et al. [13] presents the recent developments in the field of explainable AI and discusses two specific methods for the same. Singh, A. et al. [14] discuss the explainable AI methods in the medical image analysis domain. Velden, B.H.M. et al. [15] present a survey report of existing explainable AI techniques in the deep learning based medical image domain and discuss future prospects for the same. Linardatos, P. et al. [16] provides a review of various interpretability methods in the Machine Learning domain. Among the explainable AI methods, the most popular ones are LIME, Saliency Map, Occlusion Analysis, Grad-CAM, SHAP and Smooth Grad. Following are references to the research papers proposing the above methods on various datasets for interpreting CNN classification output. Ribeiro, M.T. et al. [17] proposed the LIME technique for demonstrating the explainability of the machine learning classifiers. Junkang An et al. [18] propose the LIME based explainable AI technique for interpreting the results of deep learning models using feature importance and partial dependency plots (PDPs). Alqaraawi, A. et al. [19] explores the saliency map method in detail using public datasets. Simonyan, K. et al. [20] uses two methods, one using class score and another using saliency map for analyzing the results of a deep CNN. Xiao-Hui Li et al. [21] evaluate and compare various explainable AI methods based on evaluation metrics defined. Resta, M et al. [22] provides an occlusion based technique for explanation of deep recurrent neural networks for biomedical signals. Selvaraju, R.R. et al [23] uses the Grad-CAM technique for visual explanations of the classification predictions made by deep convolutional neural networks. Cao, Q.H. et al. [24] proposes a novel explainable AI method, SeCAM (Segmentation - Class Activation Mapping) that combines the best features of LIME, CAM and the GradCAM methods for explanation of CNN prediction results. Ruigang Fu et al. [25] propose an axiom-based Grad-CAM to satisfy the axioms of sensitivity and conservation. Ioannis Kakogeorgiou et al. [26] evaluate various explainable AI methods in the context of deep learning multi-label classification for remote sensing. Lundberg, S. et al. [27] proposes the SHAP method for interpreting the model predictions using feature importance values for a particular prediction. Bach, S. et al. [28] describes an approach based on pixel-wise contributions to explain the classification predictions made by non-linear classifiers. Hooker, S. et al. [29] evaluates the performance of feature importance estimation methods used by interpretable or explainable AI methods. Ishikawa, S. et al. [30] proposes a method of explainable artificial intelligence to verify the reliability of a deep learning model for remote sensing image classification tasks. Jogani, V. et al. [31] uses various explainable AI techniques to interpret the results of CNNs for classification of lung cancer from histopathological images. Montavon, G. et al. [32] uses Deep Taylor Decomposition to explain the results of non-linear classifiers. Shrikumar, A et al. [33] proposes a method, Deep Learning Important Features (DeepLIFT) which assigns contributions towards the final output to the neurons of the deep neural network and also illustrates positive and negative contributions. Smilkov, D. et al. [34] proposes a method called SmoothGrad for explaining the results of a deep neural network. Soltani, S. et al. [35] have provided enhanced explainable AI algorithms based on cognitive theory. Springenberg, J.T. et al. [36] propose a deconvolutional approach for interpreting the classification

results of the CNNs. Sundararajan, M. et al. [37] proposed the Integrated Gradients approach for explaining the deep neural network predictions. Vermeire, T. et al. [38] propose the SEDC method which is model agnostic and is able to provide counterfactual explanations for the predictions of deep convolutional neural networks. Zeiler, M.D. et al [39] demonstrate the performance contribution of different layers of the ImageNet used for classification. Zhou, B. et al. [40] propose a modification of the global average pooling layer and implement the Class Activation Mapping (CAM) technique to improve the CNN’s performance in classification of the images without being trained specifically for the same. Wu, B. et al. [41] propose and evaluate an attention-based model for large scale image classification, and thereby explain the classification results of this approach.

3. Research Gap and Motivation

Deep learning models have proved to be accurate to classify tumors as benign or malignant and there is plenty of literature demonstrating the same. However, due to the black-box nature of deep learning models, it is difficult to interpret and accurately pinpoint the features or pixels responsible for the final model output and distinguish one class from another. There is limited literature on developing explainable AI methods to interpret and explain the classification output of these models for better data-driven decision-making in the healthcare domain. Moreover, there is a lack of adequate research work on leveraging explainable AI techniques to interpret the results of deep learning models used for classification of tumors into benign and malignant in the healthcare domain for the healthcare experts to take better data-driven decisions using AI and thus helping in early diagnosis of the disease. Hence, in this paper, we use several explainable AI techniques to explain the results of a custom ResNET60 classifier which classifies the ovarian tumors as benign and malignant with 97.50% accuracy on the test dataset and outperforms the state-of-the-art ILSVRC winning architectures.

4. Methodology

The ResNet60 model has been trained on the dataset of cropped ovarian tumor images for the classification of the tumors as benign or malignant. The training dataset comprises 3644 images with 2633 belonging to benign and 1011 belonging to the malignant classes. The model has been validated on a dataset of 1561 images comprising 1128 benign images and 433 malignant images. The model has also been tested using a test dataset of 520 images comprising 376 benign and 144 malignant images on which it achieved an accuracy of 97.50% on the test dataset. There is a class imbalance in the datasets where benign images are higher in number than malignant images. The distribution of training, validation, and testing sets is shown in Table 1. Figure 1 shows the methodology implemented. Figure 4 shows the proposed system comprising the various layers of the ResNet60 architecture, followed by the model explanations.

Table 1. Proposed ResNet60 architecture.

Stage	Layers	Conv1 Filters	Conv2 Filters	Conv3 Filters	TotalConv Filters	Stride	Padding and ReLU	Batch Norm
1	Conv	64	64	256	384	(2, 2)	Valid	Yes
1	Identity (x2)	64	64	256	384	(1, 1)	Valid	Yes
2	Conv	128	128	512	768	(2, 2)	Valid	Yes
2	Identity (x3)	128	128	512	768	(1, 1)	Valid	Yes
3	Conv	256	256	1024	1536	(2, 2)	Valid	Yes
3	Identity (x5)	256	256	1024	1536	(1, 1)	Valid	Yes
4	Conv	512	512	2048	3072	(2, 2)	Valid	Yes
4	Identity (x2)	512	512	2048	3072	(1, 1)	Valid	Yes
5	Conv	512	512	2048	3072	(1, 1)	Valid	Yes
5	Identity (x2)	512	512	2048	3072	(1, 1)	Valid	Yes
6	Conv	1024	1024	4096	6144	(2, 2)	Valid	Yes
6	Identity (x2)	1024	1024	4096	6144	(1, 1)	Valid	Yes

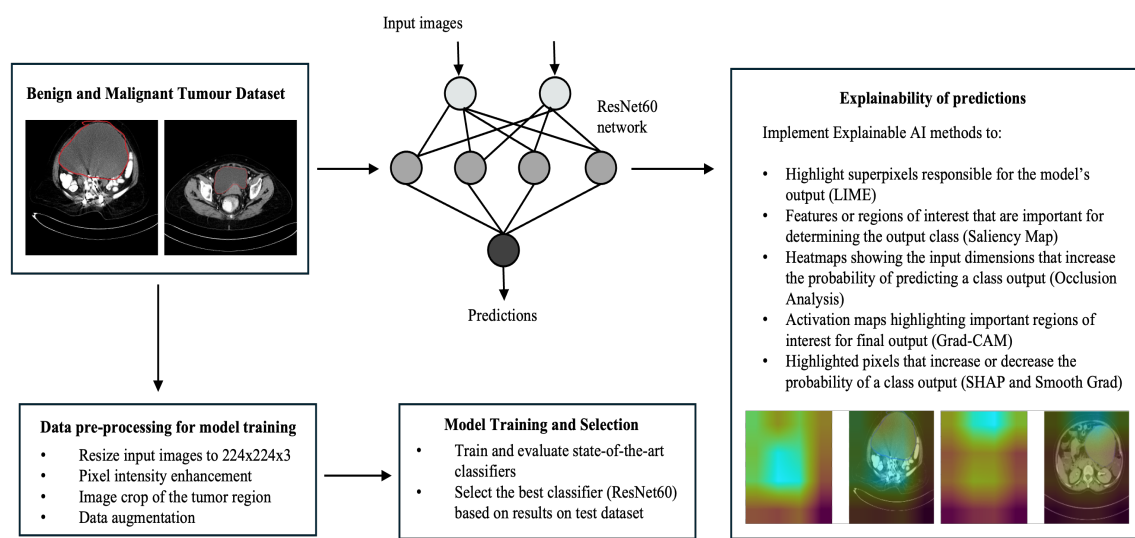


Figure 1. Methodology.

4.1. Proposed ResNet60 Architecture

The proposed architecture inputs images of a size 224x224x3. The detailed architecture is shown in Figure ??.

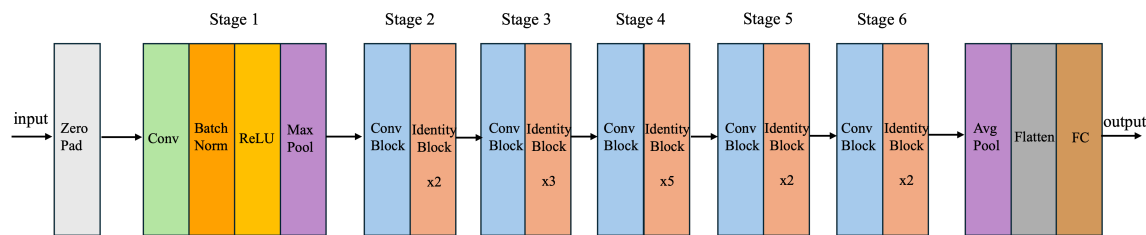


Figure 2. ResNet60 architecture for classification of ovarian tumor images into benign and malignant.

The structure of each convolutional and identity block and architecture diagram for the proposed ResNet60 architecture is displayed in Figure 3, while the complete architecture is shown in Figure 2.

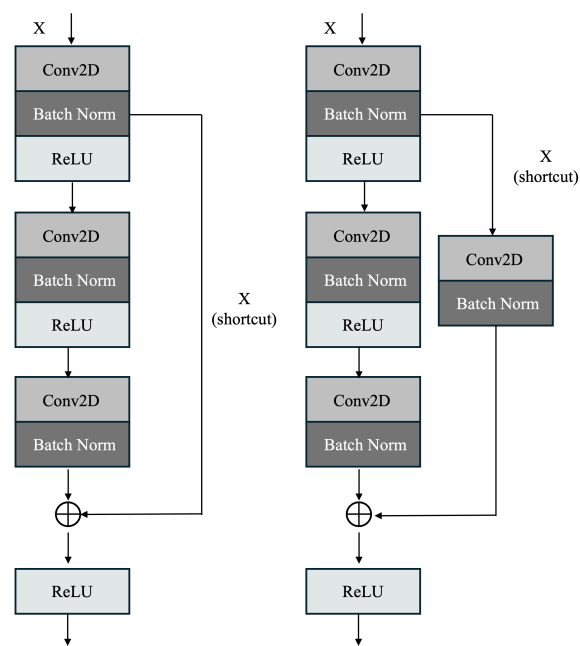


Figure 3. Inception Block (left) and Convolution Block (right).

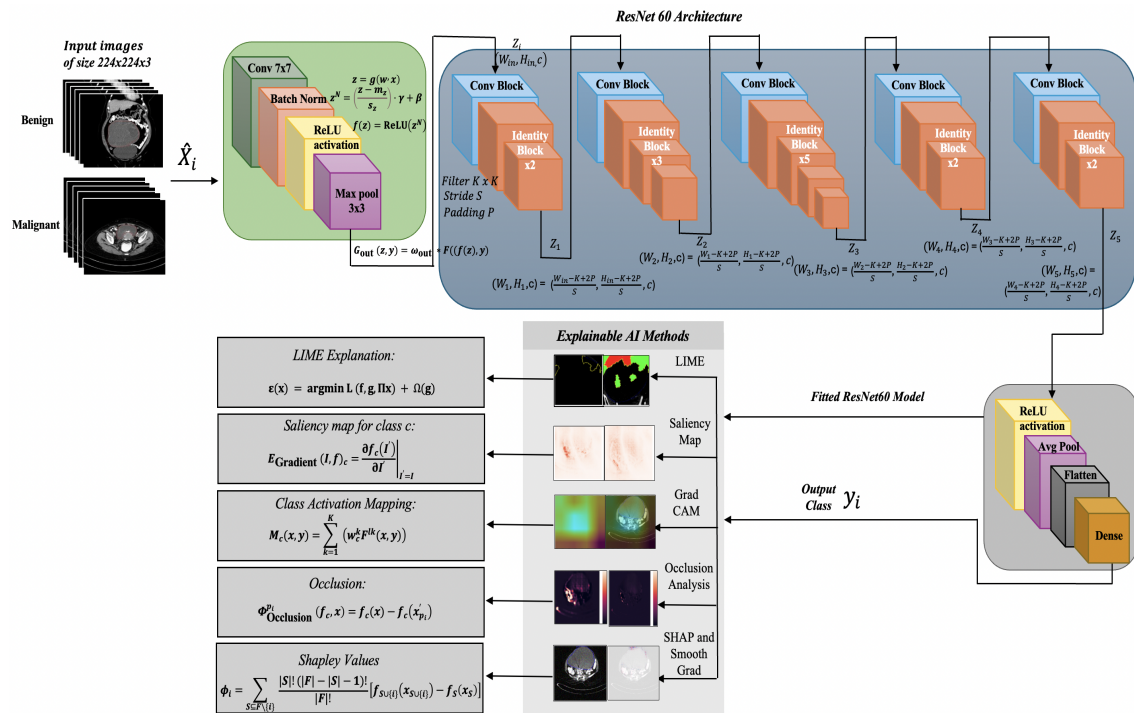


Figure 4. Proposed System.

As the deep learning architectures continued to add more layers for higher accuracy of the predicted output, the vanishing and exploding gradient problems started becoming more common with increasingly deep layers, thereby slowing down the learning and reducing the accuracy of the predicted output. This problem was solved by ResNet where residual blocks containing skip connections were introduced, which connect activations of a layer to deeper layers y skipping some layers in between.

The fundamental building blocks of the ResNet architecture and image classification is described in Algorithm 1. Figure 3 shows the basic Inception and Convolution blocks.

To interpret the classification results of the ResNet60 model, we have implemented the above-mentioned explainable AI methods. The objective of this research is to determine the important features or pixels responsible for ResNet60 to produce the final classification output. We also aimed to understand what features distinguish a benign ovarian tumor from a malignant ovarian tumor based on the CT scanned image and how the neural network captures the same.

Algorithm 1: Image Classification using custom ResNet60 Architecture**Inputs:**

Image X_i of size 512x512x3 preprocessed to $\hat{X}_i$ of size 224x224x3

Output:

Class Y_i for each image, where 0 indicates Benign and 1 indicates Malignant

ResNet Building Block:

ResNet Layer (Inputs image, number of filters, kernel size, stride, number of ResNet blocks, learning rate, epochs, padding (same), kernel initializer, kernel regularizer, activation (ReLU), batch normalization)

$$y_l = h(x_l) + F(x_l, W_l)$$

$$x_{l+1} = f(y_l)$$

For the l -th Residual Unit with K layers, x_l denotes the input feature, and $W_l = W_{l,k} |_{1 \leq k \leq K}$ denotes the set of weights and biases. The residual function is indicated by F .

Convolutional Block:

Convolutional 2D Layer (number of filters, kernel size, strides, padding, kernel initializer, kernel regularizer)

if layer = Conv2D First layer then

$$z = g(w, x); \quad z^N = \left(\frac{z - m_z}{s_z} \right) \cdot \gamma + \beta;$$

$$f(z) = \text{ReLU}(z^N)$$

$$G_{\text{out}}(z, y) = \omega_{\text{out}} * F((f(z), y))$$

For inputs x_i , z being the output of each neuron, $g()$ being the linear transformation of the neuron, w the weights, b the bias, and $f()$ the activation function, the Batch Norm produces the output z^N , with mean m_z and standard deviation s_z , and learning parameters γ and β . After applying the ReLU activation function on the Batch Norm output, a Conv2D layer is applied with kernel ω . $F((f(z), y))$ denotes the input patch of image where convolution is applied and $G_{\text{out}}(z, y)$ denotes the convolution product, y being the output of the Conv2D layer.

Bottleneck Residual Unit:

$$a(x) = \text{ReLU}(\text{BN}(x)) \quad c_k(x) = \text{Conv}_{k \times k}(a(x)) \quad \text{Bottleneck}(x) = c_1(c_3(c_1(x)))$$

The residual units comprise stacked Batch Normalization on the inputs $\text{BN}(x)$, followed by ReLU activation and convolutional layers c_3 and c_1 with filters of shape 1 and 3 respectively

$X = \text{ResNet}(\text{inputs}=I, \text{number_filters_input}, \text{kernel_size}, \text{strides}, \text{RELU activation}, \text{batch normalization}, \text{conv_first set as False})$

$X = \text{ResNet}(\text{inputs}=I, \text{number_filters_input}, \text{conv_first}=False)$

$X = \text{ResNet}(\text{inputs}=I, \text{number_filters_output}, \text{kernel_size}, \text{conv_first}=False)$

if the residual block is not the 0th then

$I = \text{ResNet}(\text{inputs}=I, \text{number_filters_output}, \text{kernel_size}, \text{strides}, \text{No activation}, \text{No batch_normalization layer added})$

ResNet Architecture:

$\text{ResNet_Layer}(\text{inputs}, \text{depth}, \text{number of classes}, \text{number_filters_input}, \text{number_residual_blocks}=\text{int}((\text{depth} - 2)/9), \text{number_stages}, \text{layer} = \text{conv_first})$

Instantiation of the residual units stack:

```
for stage in number_stages do
  for residual_block between 0 and num_residual_blocks do
    if stage is the 0th then
      number_filters_output = number_filters_input x 4
      if residual_block is the 0th (for the 1st layer and the 1st stage) then
        set activation as None no batch normalization layer to be added
      else
        set activation as ReLU
        add batch normalization layer
        set strides as 1
        number_filters_output = number_filters_input x 2
        if residual_block is the 0th (the first layer, but not the first stage) then
          set strides as 2
```

The Final Classification Layer:

$X = \text{ReLU}(\text{BN}(x))$

$$X = \text{AvgPooling}(X)_{i,j,k} = \frac{1}{f_x \cdot f_y} \sum_{m,n} X_{i-s_x+m, j-s_y+n, k}$$

$$X = \text{GlobalPooling}(X)_k = f(X[:, :, k])$$

$$O = \hat{F}_f(x, y) = I * \hat{W}_f = \sum_{x'=1}^X \left(\sum_{y'=1}^Y \left(\sum_{c=1}^C I(c, x - x', y - y') \alpha_f(c) \right) \beta_f(x') \right) \gamma_f(y')$$

$$y_{jk}(x) = f \left(\sum_{i=1}^{n_H} w_{jk} x_i + w_{j0} \right)$$

For input X , the output is defined by indices (i, j) , channel index k , horizontal and vertical strides s_x and s_y , and filter sizes f_x and f_y for the pooling window with center (i, j) . $\hat{F}_f(x, y)$ denotes the flattened layer, with the weights in the CNN being 4D filters $W \in \mathbb{R}^{C \times X \times Y \times F}$, C , number of input channels C , spacial dimensions X and Y of the filter. f denotes the index of the output channel with stride 1, with input map as $I \in \mathbb{R}^{C \times N \times M \times F}$ with spacial dimensions and channel as N , M and C respectively. $\hat{W}_f$, the unit rank filter can be expressed as:

$$\hat{W}_f = \alpha_f \times \beta_f \times \gamma_f$$

where α_f , β_f and γ_f denote the lateral, vertical and horizontal filters respectively, convolving features across channels, Y dimension and X dimension respectively. Finally, $y_{jk}(x)$ denotes the output of the final dense layer or the fully connected layer denoted by f

4.2. Explainable AI Methods

To interpret the classification results of the ResNet60 model, we have implemented the above-mentioned explainable AI methods. The objective of this research is to determine the important features or pixels responsible for ResNet60 to produce the final classification output. We also aimed to understand what features distinguish a benign ovarian tumor from a malignant ovarian tumor based on the CT scanned image and how the neural network captures the same.

LIME, which stands for Local Interpretable Model-agnostic Explanations is an Explainable AI method that aims to explain the results of a model using a local linear approach. LIME creates several artificial data points in the vicinity of the data point whose classification results it aims to explain. It uses a local linear classifier to predict the class of each of the artificially created data points. After this, the cosine similarity is calculated between each artificial data point and the original data point to determine the importance or weight of each of the artificial data points in classifying the real data point. Since the dataset consists of images, the data points are pixels, and the cosine distance is computed between the pixels to compute their weight in determining the class. Using the artificially weighted data points, a linear regression model is fitted and the features or data points with the higher coefficients in the fitted linear regression model contribute the most in determining the prediction of the model being evaluated.

Riberio et al. described a LIME-generated explanation as:

$$\varepsilon(x) = \operatorname{argmin} L(f, g, \Pi_x) + \Omega(g)$$

where g indicates the model used for approximating the function f , Π_x indicates the distance measure between z and the neighbourhood around x , z and x being instances. The complexity of explanation of model outputs using g is depicted by $\Omega(g)$.

While LIME uses feature importance to explain the importance of nearby pixels, the Saliency Map is another explainable AI method to explain the classification results of Computer Vision models. The Saliency Map highlights the region of interest comprising the pixels that are the determining factors behind the model's output. Saliency Map calculates the derivative of the class score with respect to the image and helps us identify which pixels, when changed, cause the least change in the class score.

Simonyan et al. has proposed the gradient of the output class calculated with respect to the input pixels of the image. The saliency map for a class c can be described as:

$$E_{\text{Gradient}}(I, f)_c = \left. \frac{\partial f_c(I')}{\partial I'} \right|_{I'=I}$$

$f_c(I')$ can be approximated with a linear function in the neighborhood of I by the first-order Taylor expansion: $f_c(I') \approx E(I, f)_c^I \cdot (I' - I) + b$. The weights of this approximated linear function are the generated saliency map.

Occlusion Analysis is an attribution technique that determines the importance of each feature or dimension by evaluating the model performance by having that dimension missing. The region or patch of pixels which, when dropped, causing the highest drop in the model's performance tends to have the most contribution towards the prediction results of the model.

Ioannis Kakogeorgiou et al. describe occlusion as:

$$\Phi_{\text{Occlusion}}^{p_i}(f_c, x) = f_c(x) - f_c(x'_{p_i})$$

This technique calculates the change in the approximated function $f_c(x)$ by replaced contiguous rectangular patches $p_i \in P$ of the input image with a given baseline (e.g., all-zero patch).

Grad-CAM, standing for Gradient Class Activation Map is an explainable AI technique that produces a heatmap that shows the importance of each of the classes for the model being evaluated. To achieve this, the gradient with respect to the final layer of the CNN is computed and weighed against

the output class of that layer. The heatmap is essentially a spatial map of the weight or importance of each channel towards the final output class.

Ruigang Fu et al. describe the general form of the CAM (Class Activation Mapping) as:

$$M_c(x, y) = \sum_{k=1}^K \left(w_c^k F^{lk}(x, y) \right)$$

where c denotes the class c of the input image and the CAM can be expressed as a linear combination of the features in the target layer.

SHAP (Shapley Additive Explanations) is an explainable AI method that calculates the feature importance based on Shapley values derived from cooperative game theory. The feature importance calculated using the SHAP value indicates the magnitude as well as the extent of the positive or negative effect of the feature on the final class predicted by the model, and also takes into account the presence of multicollinearity.

For calculating the importance of each feature, the model is retrained on all possible subsets of features $S \subseteq F$ from the set of all features F . The importance of each feature is thus assigned, which indicates the effect of that particular feature on the model prediction. For the $f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)$, two model are trained, one trained with the features being present, denoted by $f_{S \cup \{i\}}$ and another trained without including the features, denoted by f_S . For a set S of input features, the values are denoted by x_S . Since the effect of withdrawal of one feature depends on the subset of the other model features, the differences are calculated for all the subsets $S \subseteq F \setminus \{i\}$. Then the Shapley feature importance values are computed as a weighted average of the differences. Scott M. Lundberg describes the computation of Shapley values as:

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

SmoothGrad is an explainable AI method that is gradient-based, similar to the Saliency Map. It improves the gradient visualizations by adding a small Gaussian noise to the input image. This helps in obtaining a clean gradient visualization, free from noise, and highlights only the important pixels that are activated across all the sampled images. These important pixels determine the final classification output of the model.

Given a neural network that classifies an image into a class from a set C of classes, for each input image x , a class activation function, denoted by S_c is calculated for each class $c \in C$. The final predicted class (x) is the one with the highest score defined below (Szegedy et al., 2016; LeCun et al., 1998):

$$\text{class}(x) = \text{argmax}_{c \in C} S_c(x)$$

For locating the pixels in the image of higher importance, a sensitivity map, denoted by $M_c(x)$ is constructed as below:

$$M_c(x) = \partial S_c(x) / \partial x$$

where $M_c(x)$ is the derivative of M_c with respect to x , the input. and ∂S_c represents the derivative (i.e. gradient) of S_c .

The sensitivity maps are improved by using a Gaussian kernel to smoothen ∂S_c and taking an average of the sensitivity maps of random samples in the neighborhood of an input x . Smilkov et al. provided the below formulation for Smooth Grad:

$$\hat{M}_c(x) = \frac{1}{n} \sum_1^n M_c \left(x + \mathcal{N}(0, \sigma^2) \right)$$

where the number of samples is denoted by n , and the Gaussian noise with standard deviation σ is denoted by $\mathcal{N}(0, \sigma^2)$.

5. Results and Discussion

5.1. Data Source and Description

The dataset comprises a set of 5725 CT scan images of ovarian tumors, which have been annotated, that belong to benign and malignant classes for 53 unique patients from SDM College of Medical Sciences and Hospital located in Dharwad, Karnataka, India. The dataset contains reconstructed images as per the 3 different planes, viz. axial, sagittal, and coronal. Annotation is performed by indicating the tumors by encircling them with nearly circular-shaped polygonals. Table 2 depicts the number of instances of benign and malignant images that form the training, validation and test datasets. Figures 5 and 6 show benign and malignant tumor image samples from the dataset.

Table 2. Dataset Distribution - Number of Images.

Train		Validation		Test	
Benign	Malignant	Benign	Malignant	Benign	Malignant
2633	1011	1128	433	376	144

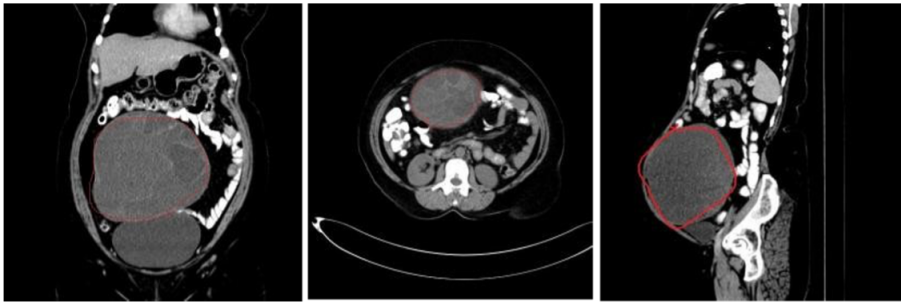


Figure 5. Benign tumors.



Figure 6. Malignant tumors.

5.2. Data Preprocessing and Dataset Preparation for Training and Evaluation

The dataset is affected by the class imbalance problem. Here, the benign sample instances are comparatively higher than the malignant ones. For training and validation, images that have been annotated were cropped according to the annotation boundary that encircle the tumors, allowing the classification models to efficiently extract and learn the features of the localized tumor. Figures 7 and 8. displays an illustration of the cropping and localization applied on annotated images:

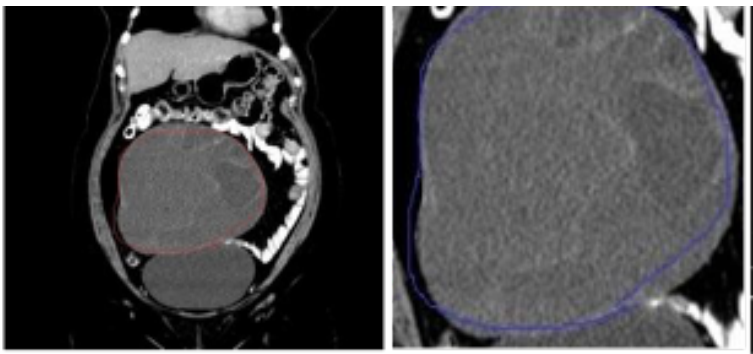


Figure 7. Original vs cropped benign tumors.

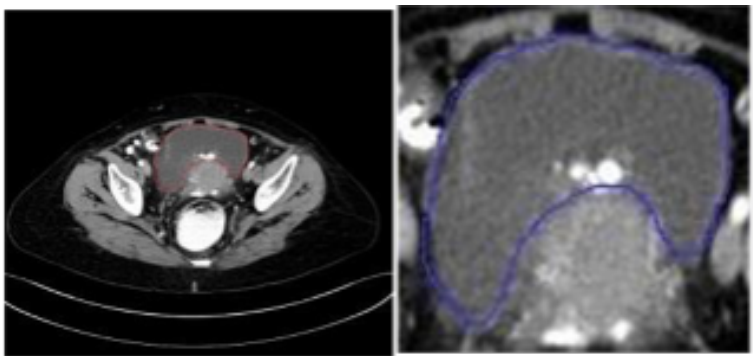


Figure 8. Original vs cropped malignant tumors.

The aforesaid architectures were trained and tested over a dataset distribution as illustrated in Table 1. The below hyperparameters were applied to the models: Input Size: 224x224x3, Optimizer: Adam, Number of epochs for training: 200 Steps per epoch: 10, Validation steps: 5, Loss: Sparse Categorical Cross Entropy (for Inception modules) and Binary Cross Entropy (for other modules)

Table 3 outlines the training and test outcomes (accuracy and loss) using 200 training epochs for various architectures. After an equal number of training epochs are passed to the model, it is evident that the suggested ResNet60 architecture exhibits a maximum accuracy for the train as well as test datasets. During validation, an accuracy of 97.50% is obtained using the given ResNet60 model. Cross-validation of train and validation samples reveals that ResNet60 typically outperforms the other designs in terms of performance. Therefore, for the given CT scan dataset, our suggested ResNet60 design fits optimally for classifying tumors as benign or malignant. It is noted that the train and validation loss incline toward convergence and sway about a constant value that is extremely near to zero as the epochs go by. The train and validation accuracy, at the start, show a rising trend and swing between 60 to 70%. Following 50 epochs, the accuracy rises steadily and varies between 80 to 90% until stabilizing at a point where the training accuracy is between 90% and 95% and the validation accuracy is between 95% and 100%. The results demonstrate that the proposed ResNet60 has steady learning while being free of bias and variance concerns since it can fit the training dataset incredibly well and generalize effectively on the validation and test datasets.

A sample of benign and malignant images of ovarian tumors from the test dataset were fed into the below-mentioned explainable AI methods for interpretation.

Table 3. Comparison of Model Performances on the CT Scan dataset for train and test.

Model Name	Variant Name	Train Accuracy	Train Loss	Test Accuracy	Test Loss
Inception	GoogLeNet (Inception v1)	91.2%	0.24	92.5%	0.25
	Inception v4	93.8%	0.12	80%	42.60
VGG	VGG 16	71.8%	2.46	72.5%	2.42
	VGG 19	74.4%	0.56	74.37%	0.57
ResNet	ResNet 50	96.2%	0.16	90%	0.52
	ResNet60 (proposed)	95%	0.19	97.5%	0.14
EfficientNet	EfficientNet B0	71.8%	0.60	72.5%	0.58
DenseNet	Densenet 121	70%	2.25	70.63%	0.54

5.3. LIME Results

The results of LIME on sample benign and malignant images are shown below in Figures 10(a) and 10(b). For interpreting the results of LIME, the cropped tumor images are fed into the method, to reduce noise from the background and only focus on the data points in the vicinity of the tumor.

From the LIME results, we see why the ResNet60 model classifies the image as benign or malignant. The yellow highlighted region indicates the superpixels where the boundary of the tumor is present. This means that the model considers the pixels on the boundary of the tumor and the neighboring pixels to determine the shape and pattern of the tumor, which helps in classifying the image as benign or malignant. These super-pixels are responsible for the final classification output result for the tumor-benign or malignant. On the right image, the area of the super-pixels colored in green indicates that these pixels cause the probability of the model predicting the predicted output class (benign or malignant) increase, while the super-pixels colored in red are the ones that decrease the probability of the model predicting the final predicted output – benign or malignant. Thus, the model can identify the special features (shape, pattern etc.) and characteristics of the tumor that are specific for the tumor to be benign or malignant, and this is highlighted by the superpixels identified by LIME, as well as the areas of the super-pixels which increase or decrease the probability of arriving at the final predicted output classification.

5.4. Saliency Map Results

The results of the saliency map as shown in Figures 11(a) and 11(b) show the highlighted pixels or the region of interest that are important in the final classification of the tumor into benign and malignant classes by the ResNet60 model. The calculation of saliency is similar to backpropagation. Here, the derivative of the class score is taken with respect to the image. This helps us identifying which pixels, if changed the least, affects the class scores the most.

5.5. Occlusion Analysis Results

From the results of the occlusion analysis as shown in Figures 12(a) and 12(b), we observe that the highlighted regions of the heatmaps generated for the benign and malignant samples indicate the most important input dimensions towards generating the final classification since the difference between the model outputs with and without those dimensions are higher compared to the other input dimensions. Below are the hyperparameters used for the occlusion analysis experiment towards the benign and malignant samples: Occluding size: 5 Occluding pixel: 0 Occluding stride: 1

We also see that the occlusion analysis method highlights the shape of the tumor and the neighboring region surrounding the tumor, hence these are important for the final predicted output.

5.6. Grad-CAM Results

From the results of Grad-CAM, as shown in Figures 13(a) and 13(b), we see the heatmaps for the images of the tumor. The size of the heatmap is determined by the spatial dimensions of the activation maps in the last convolutional layer of the network. On projecting the heatmap to the original image, we see the areas of the image that are highlighted, and these areas have been assigned the most importance for the model to predict the final output – benign or malignant for the image. It is observed that the network took the area surrounding the tumor as a determining factor for the classification of the tumor as benign and malignant. This is a reasonable explanation for the model because the neighboring areas of the tumor indicate whether the tumor has spread to the adjacent regions, which indicates the tumor is malignant, else it is benign. Also, another characteristic of the tumor considered by the model for classification as benign or malignant is the shape of the tumor, if it has a smooth shape, which has not yet spread to the neighboring areas and is concentrated in a single region, then it is a benign tumor. Otherwise, if the tumor has an irregular shape, it indicates that the tumor has disintegrated and spread into the neighboring areas, which is the characteristic of a malignant tumor. Since CNNs are feature extractors and deeper layers of the network operate in increasingly abstract spaces, the activation maps generated by the Grad-CAM method provide useful insights into which features had the most importance in determining the model's final output class. These features are not just individual pixels, but a region of interest obtained by taking activation maps of deeper convolutional layers.

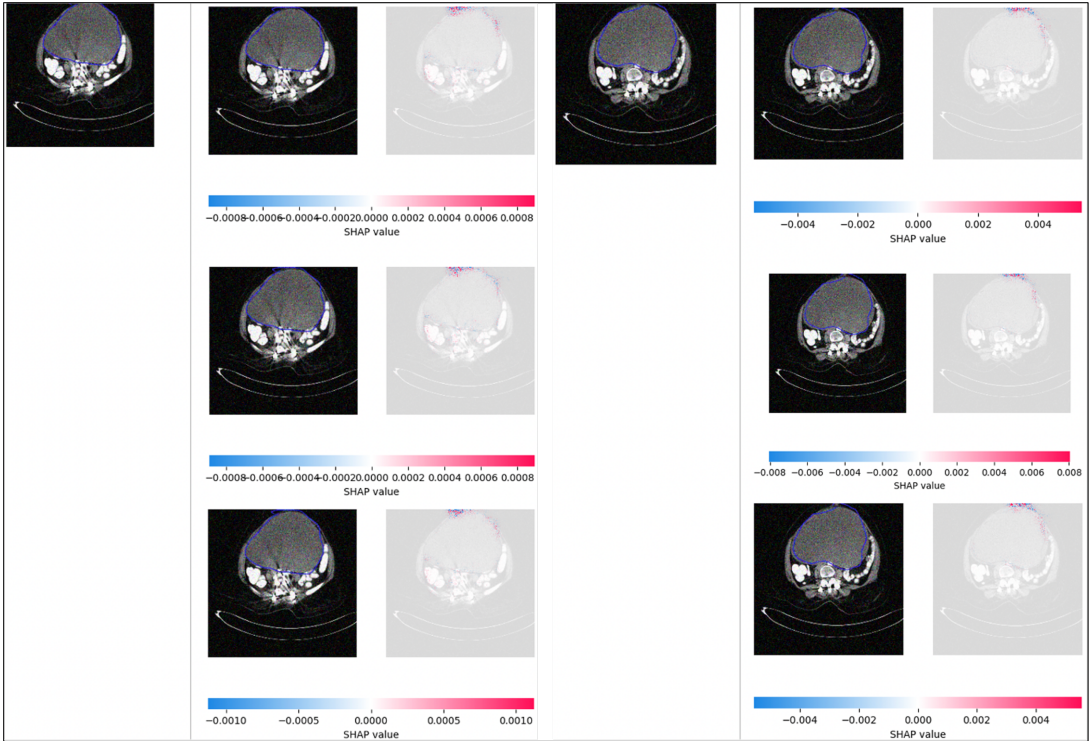
5.7. SHAP and SmoothGrad Results

The results for SHAP and SmoothGrad in Figures 9(a) and 9(b) show red-blue heatmaps on the benign and malignant images based on the Shapley values. The red colored pixels positively impact the model's final prediction output, by increasing the model's confidence for the final predicted class. The blue-colored pixels negatively impact the model's final prediction output, by decreasing the model's confidence for the final predicted class, benign or malignant as shown in the figures. We see that the red pixels are mostly concentrated around the boundary and neighboring region of the tumor, indicating that the shape and localization of the tumor, as well as the ability to spread to the neighboring regions, is taken into consideration by the ResNet60 model for the classification.

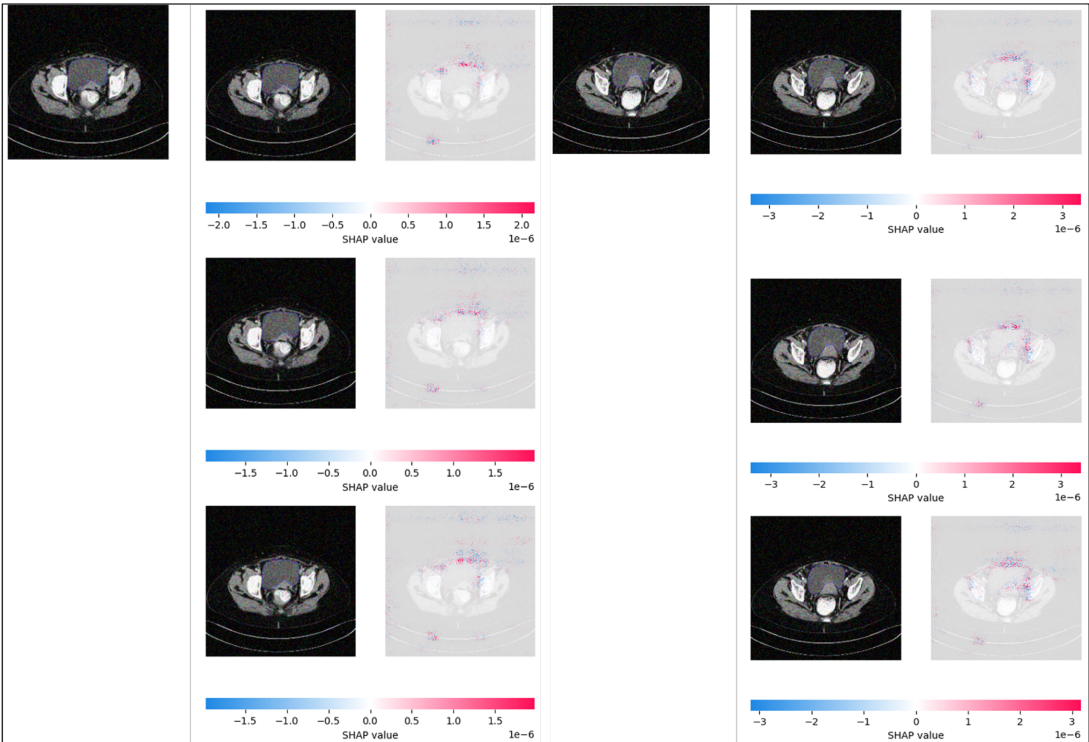
5.8. Comparison of the Explainable AI Results

It is observed that LIME can identify the superpixels responsible for the final classification output, as well as the neighboring pixels positively and negatively impacting the classification, which proves that the shape and area of concentration of the tumor is a crucial factor for the classification. The Saliency Map highlights the region of interest of the image responsible for the classification output. We observe that this region of interest is also mostly concentrated around the edge and immediate neighboring region of the tumor, which proves that these play a significant role in the final classification output. The Grad-CAM results highlight the boundary of the tumor and its adjacent regions, indicating, once again that these play a crucial role in the classification of the tumors as benign or malignant. The red-blue heatmap outputs for SHAP and SmoothGrad indicate the red pixels which increase the model's confidence towards the final predicted output and these pixels are mostly considered around the boundary and neighboring area of the tumor, thus re-affirming that the shape, localization, and ability of the tumor to spread to the neighboring areas have been given the most importance by the ResNet60 model for generating the final classification output. The heatmaps generated by the Occlusion Analysis attribution technique indicate that the shape of the tumor and neighboring region of the tumor is more important towards the final output than the other input dimensions. Although occlusion analysis is very efficient at determining the marginal effect of each input dimension if the

dimensions are independent, it has a high computational complexity since we have to evaluate the model for each of the perturbed inputs and if the input image has a higher size, the time required to generate the heatmaps increases manifold.



(a) Benign



(b) Malignant

Figure 9. Results of SHAP and Smooth Grad on benign and malignant samples of cropped ovarian tumors data. Original images (left) and highlighted regions with the SHAP and Smooth Grad results (right).

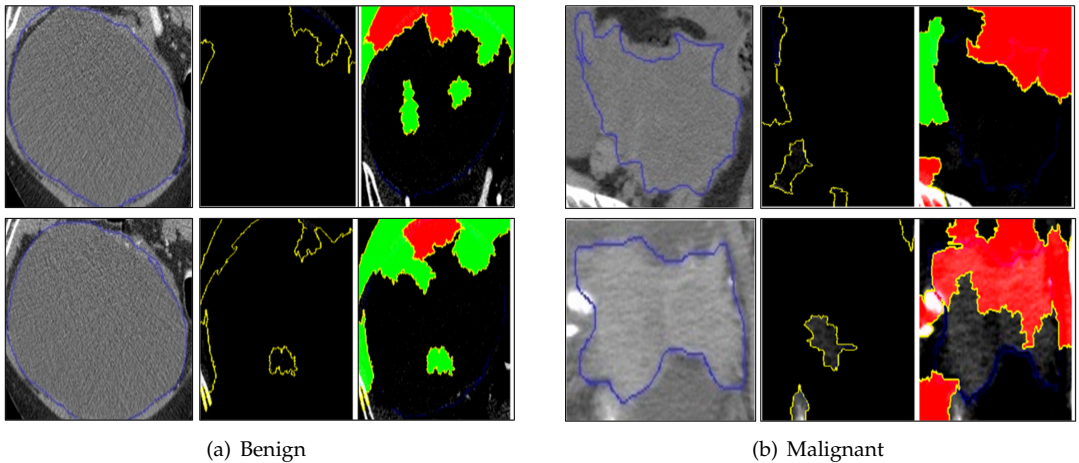


Figure 10. Results of LIME on benign and malignant samples of cropped ovarian tumors data. Original images (left) and highlighted regions with the LIME results (right).

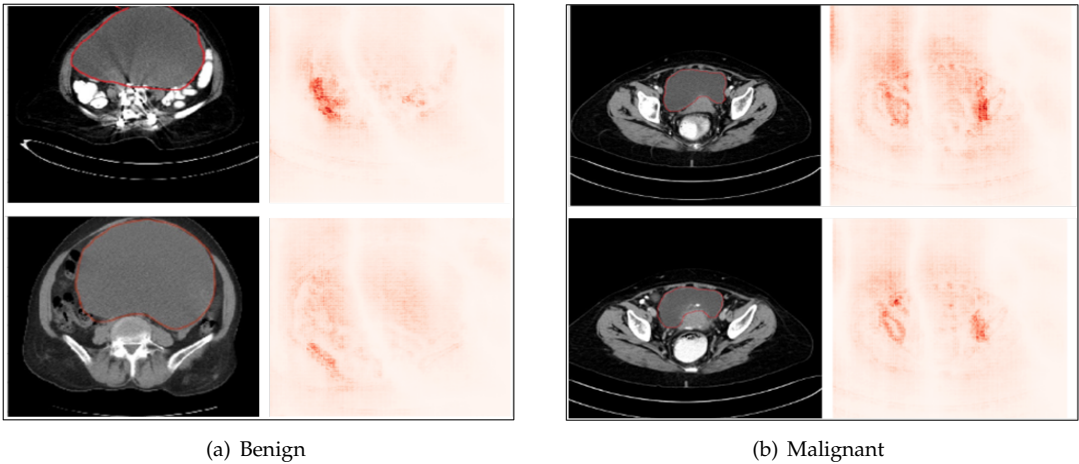


Figure 11. Results of Saliency Map on benign and malignant samples of cropped ovarian tumors data. Original images (left) and highlighted regions with the Saliency Map results (right).

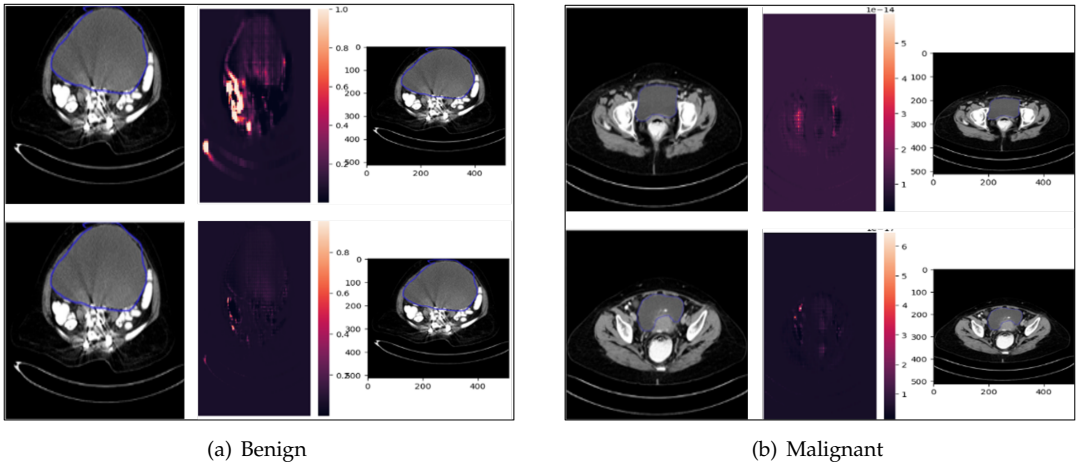


Figure 12. Results of Occlusion Analysis on benign and malignant samples of cropped ovarian tumors data. Original images (left) and highlighted regions with the Occlusion Analysis results (right).

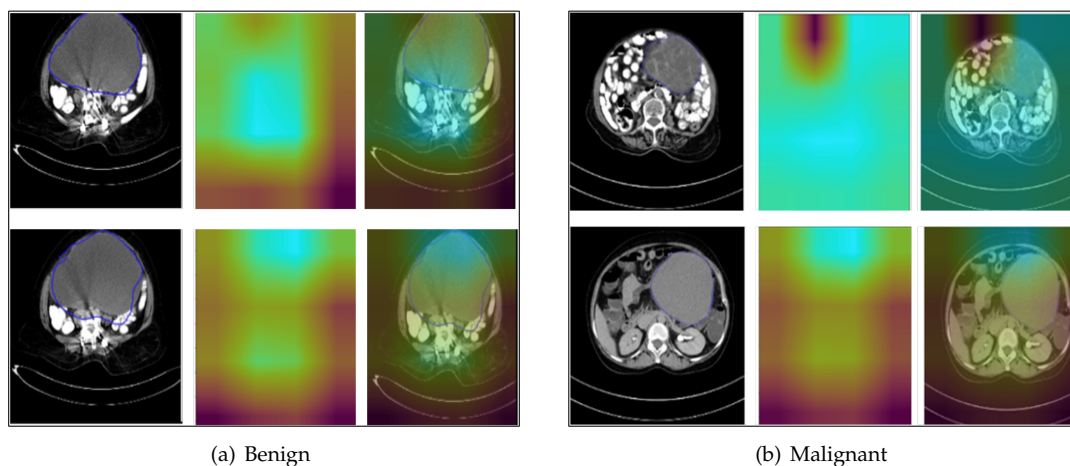


Figure 13. Results of Grad-CAM on benign and malignant samples of cropped ovarian tumors data. Original images (left) and highlighted regions with the Grad-CAM results (right).

6. Conclusion and Scope of Future Work

The leverage of AI in the healthcare domain is increasing by the day and healthcare institutions are relying on the usage of Computer vision-based classification and detection deep learning algorithms for early diagnosis of diseases like cancer, which is not intuitive to predict by human professionals. This can potentially revolutionize the healthcare and disease prediction domain and save or prolong countless lives through early diagnosis and treatment for fatal diseases like cancer. Deep learning convolutional neural network-based algorithms have proven to be one of the most accurate and efficient methods of detection of tumors from CT scans or MRI scan images. However, due to the black-box nature of these algorithms, it is difficult to interpret the results of the algorithm and understand why the model predicted a certain class. Retracing the results of the algorithm back to the individual layers can be a challenging task and it is important to understand why a neural network made a certain prediction, or, in other words, which features it relied on the most for the final predicted output. The traditional Machine Learning algorithms like Support Vector Machines and Decision Trees provide explainability, however, the problems of image classification, localization can be solved with a greater accuracy and efficiency with the deep convolutional neural networks. Sometimes due to class imbalance in the dataset, the network can be inclined towards a certain output and place importance on the wrong features. Hence, in this scenario, without some explainability methods, it would be difficult to rely on these algorithms. Class activation maps and feature importance techniques prove to be excellent in highlighting the regions of interest and thus ensuring that the algorithm results are fair and unbiased, as well as take into consideration the right features for prediction. This is crucial for healthcare practitioners to continue to rely on and have faith in deep learning algorithms to help in their decision-making processes. Thus, to interpret the results of these models, explainable AI methods are being implemented. Inspecting intermediate features at each layer is one of the popular methods to interpret model representation at various layers. The second approach is attention based, which highlights the input features that are given most importance during prediction. Explainable AI methods, on the other hand, help to analyze the decision making process of neural networks with the help of class activation maps and local explanations, and have been implemented in this paper. In some binary or multiclass classification problems, there is also a scope of creating an ensemble of neural networks combined with decision tree or rule based approaches to enhance the interpretability of predictions. Finally, introducing dropout or regularization in networks, and thus increasing the sparsity of the networks can help in identifying the features that are most important towards the final predicted output.

In this paper, we have implemented several state-of-the-art techniques for classification of tumors in the ovary as benign or malignant and observed that a custom ResNet60 architecture yields best

results on the test dataset. Thus, we have evaluated and interpreted the results of a custom novel ResNet60 classifier on a dataset of benign and malignant ovarian tumor images using different types of explainable AI techniques. Since the model has 97.5% accuracy of the test dataset, the results of the explainable AI techniques show that the ResNet60 model has given importance to the correct features and regions of interest in determining whether an image of an ovarian tumor is a benign or malignant one. The results of LIME showed that the model has given importance to the boundary and shape of the tumor, which is an important factor to consider for the classification of benign from malignant images. Results of the saliency map show the region of interest that the model has focused on the most for the generation of the output. The Grad-CAM results show that the model has given the most importance to the neighboring areas of the tumor as well as the shape and boundary of the tumor for distinguishing a benign tumor from a malignant one. Thus, with the help of these explainable AI methods, we interpreted the results of the ResNet60 model and understood which features or regions of interest it has given importance for generating the final classification output.

There is a lot of scope in utilizing gradient-based approaches and feature-importance techniques for explaining deep neural networks better. As a next step, the deconvolution and guided backpropagation techniques can be implemented on the dataset of ovarian images for better interpretation. PatternNet, a newly emerging deep neural network architecture is being used for identifying and visualizing patterns in the image data which can explain the classification results of convolutional neural networks on these images. Mining visual patterns in images have the potential to reveal useful information about the intrinsic properties of objects and images, which influence neural networks to arrive at a certain classification decision.

The source code for CNN architectures and Explainable AI methods in this paper are available at: [Explainable AI - Ovarian Tumor Classification \(GitHub\)](#).

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in collaboration with SDM College of Medical Sciences and Hospitals, and approved by the Hospital Ethical Approval Committee on 18 March 2021, approval code: Ref: SDMCMS&H/Principal Office/061/2021.

Informed Consent Statement: The work is carried out in collaboration with SDM College of Medical Sciences and Hospital. The dataset is obtained from the hospital with the proper ethical approvals from the hospital authority.

Data Availability Statement: The dataset is available on request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wojtyła, C., Bertuccio, P., Giermaziak, W., Santucci, C., Odone, A., Ciebiera, M., ... & La Vecchia, C. (2023). European trends in ovarian cancer mortality, 1990–2020 and predictions to 2025. *European Journal of Cancer*, 194, 113350.
2. Asangba, A. E., Chen, J., Goergen, K. M., Larson, M. C., Oberg, A. L., Casarin, J., ... & Walther-Antonio, M. R. (2023). Diagnostic and prognostic potential of the microbiome in ovarian cancer treatment response. *Scientific reports*, 13(1), 730.
3. Jan, Y. T., Tsai, P. S., Huang, W. H., Chou, L. Y., Huang, S. C., Wang, J. Z., ... & Wu, T. H. (2023). Machine learning combined with radiomics and deep learning features extracted from CT images: a novel AI model to distinguish benign from malignant ovarian tumors. *Insights into Imaging*, 14(1), 68.
4. Vela-Vallespín, C., Medina-Perucha, L., Jacques-Aviñó, C., Codern-Bové, N., Harris, M., Borrás, J. M., & Marzo-Castillejo, M. (2023). Women's experiences along the ovarian cancer diagnostic pathway in Catalonia: A qualitative study. *Health Expectations*, 26(1), 476-487.
5. Zacharias, J., von Zahn, M., Chen, J., & Hinz, O. (2022). Designing a feature selection method based on explainable artificial intelligence. *Electronic Markets*, 32(4), 2159-2184.
6. Gupta, M., Sharma, S. K., & Sampada, G. C. (2023). Classification of Brain Tumor Images Using CNN. *Computational Intelligence and Neuroscience*, 2023.

7. Yadav, S. S., & Jadhav, S. M. (2019). Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big data*, 6(1), 1-18.
8. Liang, J. (2020, September). Image classification based on RESNET. *Journal of Physics Conference Series* 1634(1)
9. Saranya, A., & Subhashini, R. (2023). A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decision analytics journal*, 100230.
10. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K. R. (2010). How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11, 1803-1831.
11. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II* 8 (pp. 563-574). Springer International Publishing.
12. Yang, W., Wei, Y., Wei, H., Chen, Y., Huang, G., Li, X., ... & Kang, B. (2023). Survey on explainable AI: From approaches, limitations and Applications aspects. *Human-Centric Intelligent Systems*, 3(3), 161-188.
13. Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.
14. Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *Journal of imaging*, 6(6), 52.
15. Van der Velden, B. H., Kuijff, H. J., Gilhuijs, K. G., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 79, 102470.
16. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1), 18.
17. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
18. An, J., Zhang, Y., & Joe, I. (2023). Specific-Input LIME Explanations for Tabular Data Based on Deep Learning Models. *Applied Sciences*, 13(15), 8782.
19. Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., & Berthouze, N. (2020, March). Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th international conference on intelligent user interfaces* (pp. 275-285).
20. Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
21. Li, X. H., Shi, Y., Li, H., Bai, W., Song, Y., Cao, C. C., & Chen, L. (2020). Quantitative evaluations on saliency methods: An experimental study. *arXiv preprint arXiv:2012.15616*.
22. Resta, M., Monreale, A., & Bacciu, D. (2021). Occlusion-based explanations in deep recurrent models for biomedical signals. *Entropy*, 23(8), 1064.
23. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).
24. Cao, Q. H., Nguyen, T. T. H., Nguyen, V. T. K., & Nguyen, X. P. (2023). A Novel Explainable Artificial Intelligence Model in Image Classification problem. *arXiv preprint arXiv:2307.04137*.
25. Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., & Li, B. (2020). Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv preprint arXiv:2008.02312*.
26. Kakogeorgiou, I., & Karantza, K. (2021). Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 103, 102520.
27. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
28. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), e0130140.
29. Hooker, S., Erhan, D., Kindermans, P. J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32.

30. Ishikawa, S. N., Todo, M., Taki, M., Uchiyama, Y., Matsunaga, K., Lin, P., ... & Yasui, M. (2023). Example-based explainable AI and its application for remote sensing image classification. *International Journal of Applied Earth Observation and Geoinformation*, 118, 103215.
31. Shivhare, I., Jogani, V., Purohit, J., & Shrawne, S. C. (2023, January). Analysis of Explainable Artificial Intelligence Methods on Medical Image Classification. In *2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)* (pp. 1-5). IEEE.
32. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65, 211-222.
33. Shrikumar, A., Greenside, P., & Kundaje, A. (2017, July). Learning important features through propagating activation differences. In *International conference on machine learning* (pp. 3145-3153). PMLR.
34. Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.
35. Soltani, S., Kaufman, R. A., & Pazzani, M. J. (2022). User-centric enhancements to explainable ai algorithms for image classification. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44).
36. Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2015). Striving for simplicity: The all convolutional net. In *arxiv: cs. arXiv preprint arXiv:1412.6806*.
37. Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319-3328). PMLR.
38. Vermeire, T., Brughmans, D., Goethals, S., de Oliveira, R. M. B., & Martens, D. (2022). Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications*, 25(2), 315-335.
39. Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I* 13 (pp. 818-833). Springer International Publishing.
40. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921-2929).
41. Wu, B., Fan, Y. and Mao, L. (2021) Large-scale image classification with explainable deep learning scheme, *Research Square*.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.