Universiteit Leiden
The Netherlands

BSc Data Science and Artificial Intelligence

Fractal Analysis as a Quantitative Explainability Tool for CNN Decision Making in Breast Cancer Cell Classification

Kaier Hartman

Supervisors:
Dr. Lu Cao

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

15/01/26

Abstract

This study aimed to investigate the viability of fractal dimension as an explainability tool for the decision-making process of convolutional neural networks (CNNs), in particular for the classification of breast cancer using the publicly available BreakHis dataset. Neural networks such as CNNs are capable of achieving high accuracies in image classification tasks, including distinguishing malignant and benign cells. As a result, this technique is being rolled out in the medical field as an assistance tool for experts to aid their decision-making. The technique, however, is largely a black box, lacking insight into the reasoning the network uses to reach its classification. Currently available explainability tools for breast cancer detection suffer from the fact that the produced attention heat maps still require expert knowledge for interpretation.

Fractal dimension, as a measure of roughness, can potentially serve as an easy-to-interpret value for checking CNN outputs. To investigate this, a pre-trained ResNet-50 model was fine-tuned on the BreakHis dataset to an accuracy of 83.46%. Based on this model, saliency maps were created using the HiResCAM method, from which CNN focus areas were obtained. Finally, fractal dimension, and several other measures for comparison, were computed to compare focus and non-focus areas of malignant and benign images to analyze their discriminative power. The experiment demonstrated that for 40x zoom breast cancer images, fractal dimension showed strong potential by having significantly different values in the CNN focus areas for malignant images compared to non-focus areas and benign images. For the 400x zoom dataset and the other more standard measures, this was not the case. The results require further research to test generalizability, but support the possibility of fractal dimension being a useful and easy-to-understand measure for providing insight into CNN decision-making within the field of breast cancer research.

Contents

1 Introduction **1**

2 The Core Concepts **2**

2.1 Convolutional Neural Networks 2

2.2 Saliency Maps 2

2.3 Fractals 3

2.4 Fractal Analysis 6

3 Overview and Problem Definition **8**

3.1 The Current Situation 8

3.2 The Role of Fractal Analysis 9

4 Methodology **10**

4.1 The Dataset 10

4.2 The CNN 13

4.3 Saliency Maps 14

4.4 Masking 15

4.5 Pattern analysis 15

4.6 Statistical Tests 16

5 Results and Interpretation **17**

5.1 Results 17

5.2 Interpretation 19

6 Discussion **20**

7 Conclusion **22**

References **27**

A Code and Acknowledgments **27**

B Full Results Tables **27**

1 Introduction

Breast cancer is the second most common form of cancer according to the World Cancer Research Fund, and by far the most common form found in women [Wor24]. Due to timely detection and improved treatment, the mortality rate has been steadily decreasing over the years. A method that is being used in recent years to help detect the illness in early stages is artificial intelligence (AI), in the form of convolutional neural networks (CNNs). CNNs can train on large amounts of preprocessed annotated microscope images of cell tissue to be able to classify new cases with accuracies near 95% [AYM22, EBI24]. A downside of this technique, however, is the black box nature of CNNs. Based on its training, the neural network produces a classification, with limited insight into the underlying decision-making process. This exact understanding, however, is extremely important for building trust, ensuring safe deployment and being able to learn from mistakes, especially in contexts such as the medical field [ABV+20].

The research area known as explainable artificial intelligence (XAI) focuses on making AI decision-making more transparent. From this field, the technique commonly applied to CNNs is the creation of saliency maps, for example applicable to plant disease identification or COVID-19 prediction [NJS+18, MCAS+22]. Saliency maps are heatmaps, which show which parts of the input image were most important for the CNN to reach its classification decision. For many applications, this gives valuable insight into the neural network's process by whether the indicated focus area is appropriate for the given task [SCD+17, GHY+23]. Within the context of breast cancer classification, however, the generated heatmaps require expert knowledge for analysis. Given how important understanding within this context is for everyone involved [ABV+20], this is a gap to be bridged. In this project, the goal is to test a method that expands on standard saliency maps to provide explainability for the CNN decision-making process involving breast cancer detection images.

The method that is being put to the test is fractal analysis. Fractal analysis computes different measures derived from fractal patterns, which are also known as self-repeating patterns [Man83]. The measure of interest specifically is fractal dimension, which can be summarized as a measure of roughness[Man06]. This appears to be an appropriate tool, as malignant cancer cells are generally more irregular and complex compared to their benign counterparts. In previous work, fractal analysis has already shown success in brain health classification [KUA+25] and Alzheimer's detection [KWC+24]. Moreover, a combination of fractal analysis and CNNs has shown improvement over CNNs by themselves on histology images [RLNZdN22], supporting the possible usefulness of this measure. By first training a CNN on the BreakHis Dataset of breast cell tissue images, after that extracting the focus areas from the saliency maps, and finally comparing the fractal dimension within CNN focus and non-focus areas of benign and malignant cell images using statistical tests, the following research question is answered:

Can fractal analysis provide effective quantitative measures to enhance interpretability of CNN decision-making in breast cancer histopathological image classification?

Or in other words, is fractal analysis a suitable tool to help explain the decision-making process of CNNs, in particular for breast cancer detection images.

To be able to answer this question, this report is structured as follows; First in Section 2 the different

relevant concepts for the project are introduced and briefly explained to give an understanding of the different subjects involved. After that, in Section 3 an overview of related research is given, showing the basis for the use of fractal analysis within this field, as well as showing the current situation. In Section 4 the methodology of the experiment is described, followed by the results in Section 5. Finally in Section 6 the implications, points of interest and angles for further research are set out. In the end, the goal is to be able to answer whether fractal dimension is a suitable and comprehensible measure for enhancing explainability of CNNs when used for breast cancer detection tasks.

2 The Core Concepts

In this section, the major concepts present in this report are introduced. First, a short description of convolutional neural networks (CNNs) is given. This is followed by an explanation of saliency maps and the different implementations considered. Next, the concept of fractals is explained. Finally, the measures commonly used in fractal analysis are discussed.

2.1 Convolutional Neural Networks

A convolutional neural network (CNN) is a form of artificial intelligence (AI) that falls under the category of deep learning models. A neural network consists of layers of neurons, which are computational units with trainable weights. During training, these weights are tuned based on large amounts of data with the help of feedback signals. The term deep learning refers to the multi-layer structure of the network, in which different layers learn features of increasing complexity from the input data.

CNNs are neural networks that contain special types of layers, including convolutional layers. Convolutional layers are used for the detection of patterns such as edges, textures and shapes. These layers in combination with the use of shared weights and having local connectivity, make CNNs especially effective for image-based tasks [ZRSC15]. Common applications of CNNs include image classification, image segmentation, and object detection. Because CNNs can handle large amounts of data and automatically learn features without the need for manual feature engineering, they have shown great potential within the medical field [HC24, VTA+19, NRN25].

2.2 Saliency Maps

Saliency maps are a technique from the field of Explainable Artificial Intelligence (XAI) that aims to provide insight into the decision-making process of CNNs. They are typically visualized as heatmaps highlighting areas of an input image that are most influential on the predicted class [SVZ14].

One of the most widely used saliency map methods for CNNs is Gradient-weighted Class Activation Mapping (Grad-CAM) [SCD+17]. Grad-CAM generates the heatmaps by computing the gradient of the class score based on one of the final layers of the network. In these layers, the highest complexity

features are handled, making them the most informative for the generation of the saliency maps.

Throughout this project the HiRes-CAM method was used, as it extends the Grad-CAM method by using higher resolution gradient information. This results in more precise and detailed saliency maps, at the cost of higher computational complexity [DC20]. Figure 1 shows an example of a saliency map generated for an image of a cat using the HiResCAM method.

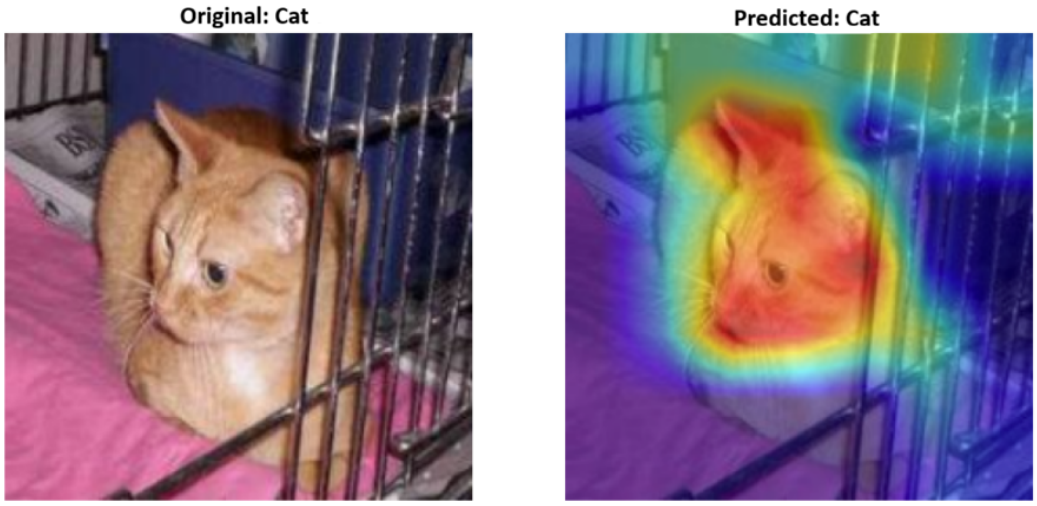


Figure 1: HiResCAM generated heatmap on a picture of a cat (right) and the original image (left).

2.3 Fractals

The term *fractal* was coined by mathematician Benoit Mandelbrot in 1975 [Man83], formalizing ideas that had appeared over the previous centuries in the work of other mathematicians. These patterns are often characterized as self-similar, showing similar shapes or repeated patterns upon zooming in. Following the formalization by Mandelbrot, the mathematical field of fractal geometry gained popularity and expanded over time. Outside of mathematics, fractal patterns are often associated with beauty, as they frequently appear in nature and art. Examples from nature include, among others, tree branches and coastlines. Additional examples are shown in Figure 2.

Working towards the definition of fractals as introduced by Mandelbrot, several classic mathematical examples will be presented. In 1904, Helge von Koch introduced the Koch curve, based on which Edward Kasner introduced the Koch snowflake. See Figure 3 for the construction of this pattern. It starts with a simple equilateral triangle. Then for each iteration, every edge is divided into three equal parts, of which the middle part is replaced by two outward pointing edges, forming a new equilateral triangle. Repeating this process infinitely many times, creates the Koch snowflake, as found in Figure 4. By focusing on the edges of the figure, the self-similarity revealed by zooming in can be observed.

Another classic example is the Sierpiński triangle, named after Polish mathematician Wacław Sierpiński. Its construction starts with a solid equilateral triangle. On the first iteration, the triangle is changed into four smaller equilateral triangles, three solid ones around the edges and an empty

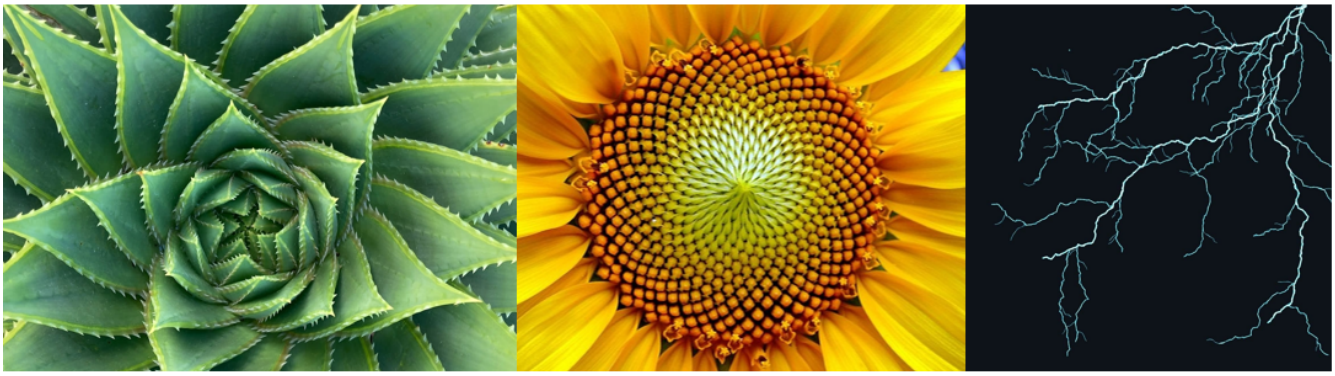


Figure 2: Examples of fractals in nature.

triangle upside down in the middle. Repeating this process for every available solid triangle for each iteration creates the figure as seen in Figure 4. As in the previous example, the self-similarity characteristic can be observed from this figure.

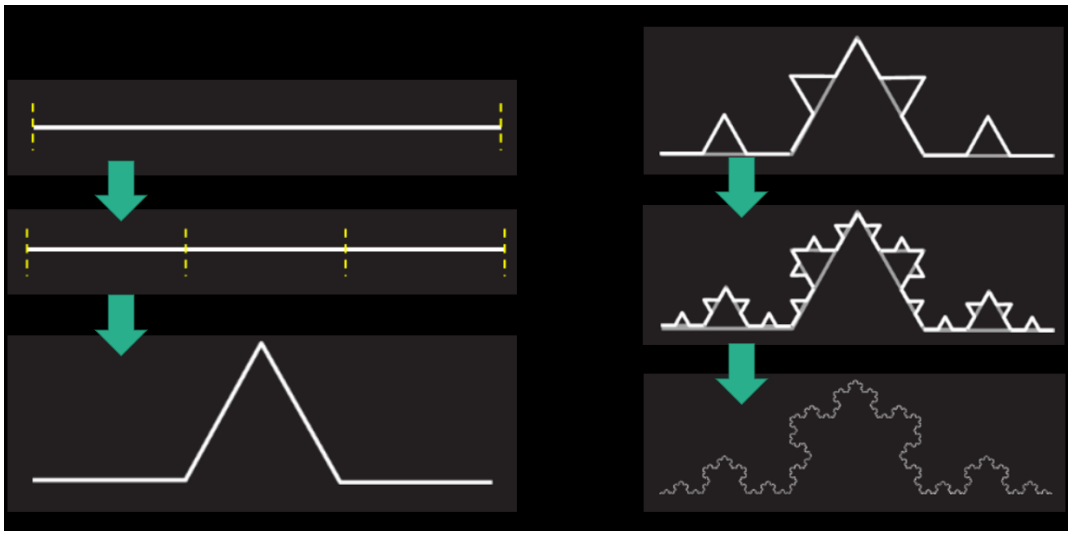


Figure 3: A step by step visual of the construction of the Koch curve (a side of the Koch snowflake).

Within nature, for example for a tree branch, the self-similarity is not perfect or infinite, as was the case for the Koch snowflake or the Sierpiński triangle. Because of this, perfect self-similarity is not suited as the exact definition of a fractal. To get to a proper definition, an overview of the concept of dimension is required.

In Figure 6 a visual overview of dimension as a function of scale factor and number of self-similar pieces can be seen. For example, if we have a one-dimensional line and scale it by a scale factor of two, the line becomes two times as long, or in other words, we now have two self-similar pieces of the original shape. Similarly, for a two-dimensional square scaling each edge by a factor of two, leads to a square with four times the area or four self-similar pieces. Applying a scaling factor of two to a three-dimensional cube this time, gives a cube with eight times the volume, or eight self-similar pieces. This relationship can be summarized as $N = R^D$ and rewritten as $D = log(N)/log(R)$, where D is dimension, R is the scaling factor and N is the resulting number of self-similar pieces.

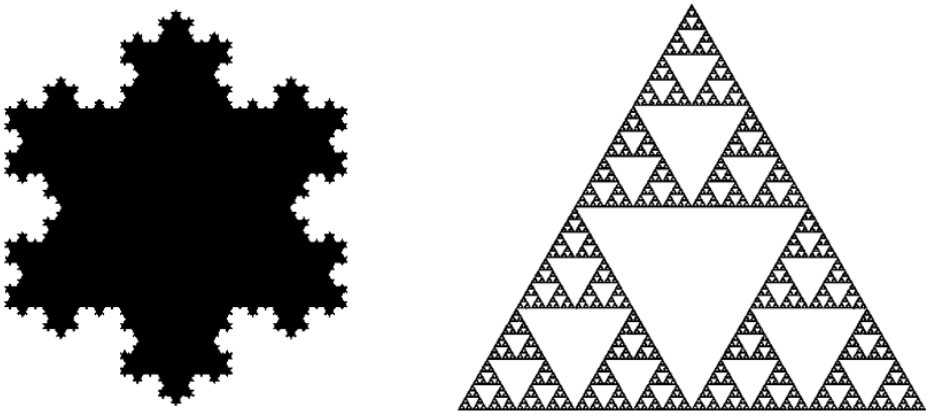


Figure 4: The Koch snowflake (left) and the Sierpiński triangle (right).

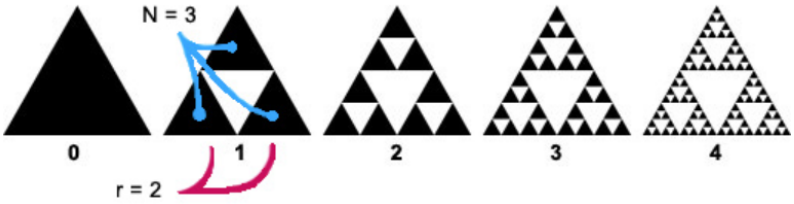


Figure 5: Scaling factor and number of self-similar pieces from the construction of the Sierpiński triangle.

Computing dimension using this formula is known as the Hausdorff Dimension and can be applied to the fractals that have been discussed previously, such as the Sierpiński triangle. Scaling the figure by a factor of two along the edges creates a larger version of the figure containing three self-similar parts compared to the original, see also Figure 5. Plugging these numbers into the derived Hausdorff dimension formula gives $D = log(3)/log(2) \approx 1.585$. The same process can be applied to the Koch snowflake by focusing on the construction of the Koch curve. In the bottom left of Figure 3, each subpart is 1/3 the length of the original edge. This means that the scaling factor used here is three, while it consists of four self-similar pieces, leading to a Hausdorff dimension of $D = log(4)/log(3) \approx 1.262$. The non-integer nature of the computed dimensions might stand out, which is a sufficient condition to qualify as a fractal pattern. However, for a definition applicable to all sorts of patterns, also those out in nature, Mandelbrot defined a fractal as a pattern with a Hausdorff dimension strictly greater than its topological dimension [Man83]. For the Sierpiński Triangle and the Koch snowflake the topological dimension is equal to 1, making them fractal under this definition. Non-integer dimension being a sufficient condition follows from the Hausdorff dimension always being larger or equal to the topological dimension for any shape. For the above calculation of the Hausdorff dimension, the perfect self-similarity characteristic of the patterns was used. To be able to apply the same definition to patterns in nature, an approximation method exists, which will be described in the next section.

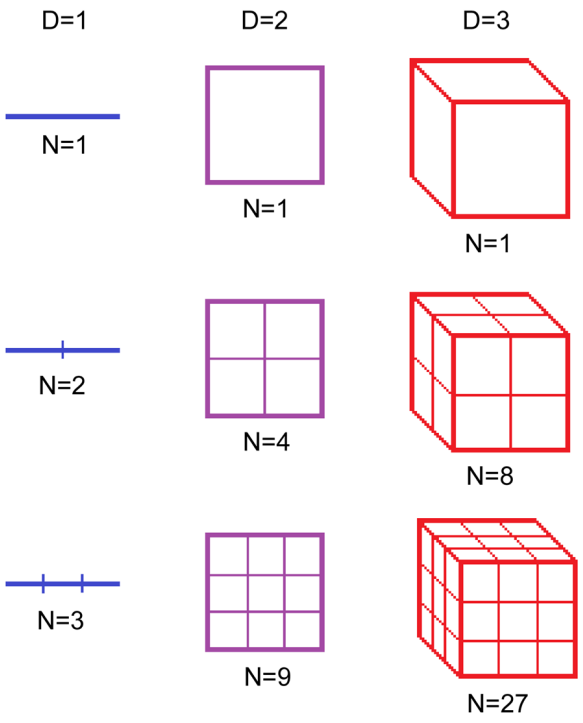


Figure 6: Visual overview of the relationship between dimension, scaling factor and number of self-similar pieces.

2.4 Fractal Analysis

Fractal analysis aims to compute fractal features from patterns. One of these features is an approximation of the Hausdorff dimension, often referred to as the fractal dimension. The method to achieve this is called the box-counting method of which a variation will be used later in the experiment. The box-counting method works by putting the pattern of interest on a grid and counting how many boxes the pattern touches. After that, the pattern is scaled by an arbitrary factor and the number of boxes touched is counted again. This process can be repeated several times to calculate the slope of the relationship between scale and number of boxes [FPDS99]. This value is the fractal dimension and approximates the Hausdorff dimension. In Figure 7 this process is visualized using the coast of Great-Britain, as coastlines are a famous example of real life fractals. For the coast of Great-Britain the fractal dimension is approximately 1.21. Using the same technique, the coast of Norway can be approximated as 1.52. The difference in fractal dimension can be seen visually by comparing the two coastlines, as is done in Figure 8. From this figure, it can be seen that the coast of Norway is more complex, containing more curves and crevices in comparison to its British counterpart. This observation is reflected in the higher fractal dimension, thus showing that fractal dimension can also be characterized as a measure of roughness.

Another fractal feature of later use is lacunarity. Lacunarity is often understood as a measure of gappiness, that is to say, the variability in size, number, and positioning of the gaps in a figure [Dal00]. Think of two forests with the same amount of trees. The first forest has all trees evenly spread, meaning it has uniform gaps and therefore low lacunarity. The second forest has trees

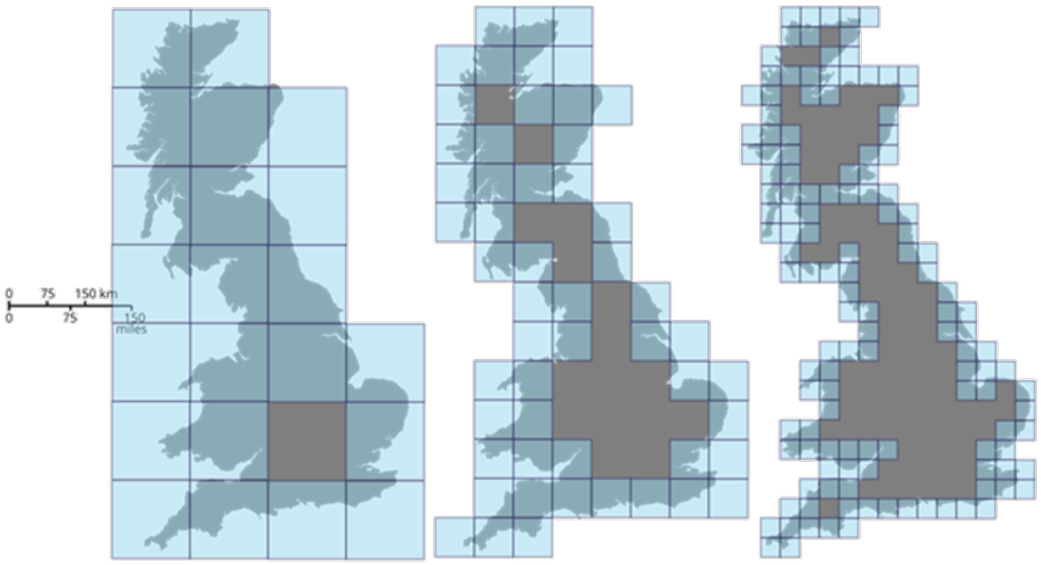


Figure 7: Visualization of the box-counting method applied to the coastline of Great Britain.

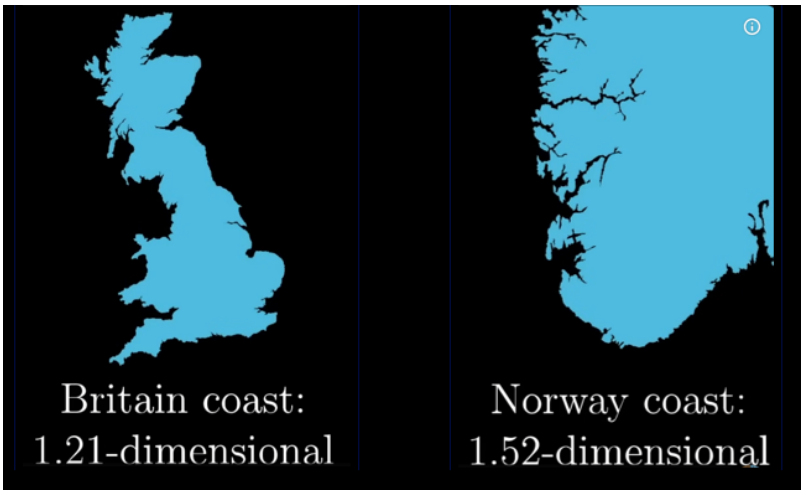


Figure 8: Comparison of the British coast and the Norwegian coast.

randomly spread, with open fields and dense areas. This layout leads to higher variability of gaps and thus a higher lacunarity. See Figure 9 for an additional visual example.

These are the concepts that will be of use later during the experiment. In the next section an overview is given of the current state of the field and where fractal analysis fits in.

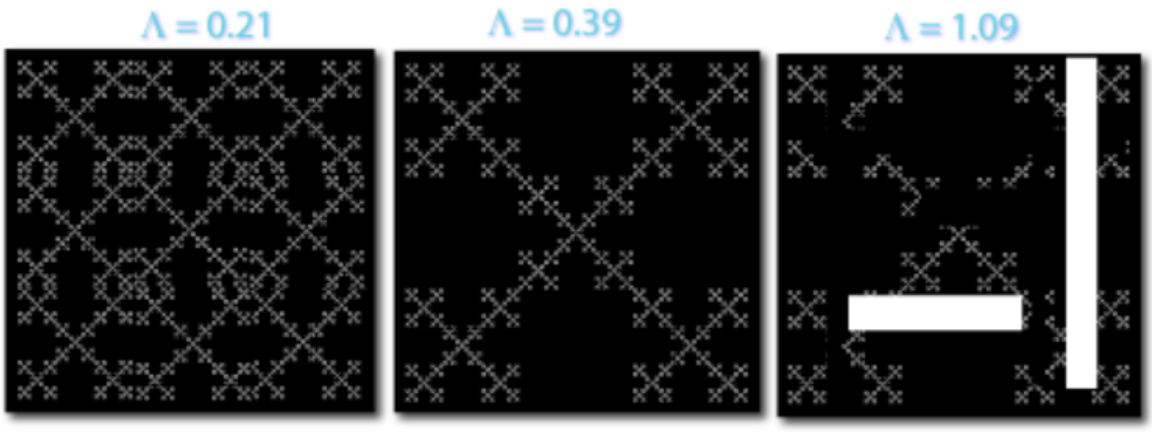


Figure 9: Visual overview of a low, medium, and high lacunarity pattern from left to right.

3 Overview and Problem Definition

3.1 The Current Situation

CNNs in the medical field are becoming more and more common and are beginning to be used in a wide range of domains. For example, CNNs can contribute to the diagnosis of brain tumors [KRM+24] and the detection of pneumonia [VTA+19]. They are also already being used for the early recognition of sepsis [HC24]. This also applies to the domain of this study, the detection of breast cancer [KRM+24], with reported accuracies between 89% and 98% according to Nasir et al. [NRN25]. The high classification accuracy of CNNs contributes to earlier and faster detection of malignancies, leading to faster and better care.

Since the technology is in an introductory stage, their main use as of now is as an assistance tool for professionals, to aid their decision-making. Before the next step can be taken, by having a CNN-type technique with more agency, the problem of the black-box nature of deep learning models needs to be mitigated. Especially for decisions as important as those in the medical field, understanding the reasoning behind the decision is of crucial importance. This insight is vital, not only for people's trust, but also for guaranteeing safe usage and enabling feedback loops for improvement [ABV+20].

The saliency map method as described in Section 2.2 is the method of choice for CNNs. By looking back on Figure 1, showing this technique applied to an example image of a cat, it can be seen that this technique rather accurately captures an area of the image that is expected to be the focus for a classification task. Having this visual might increase one's confidence in the CNN output, as well as increase the trust that the right reasoning was used to come to this decision. In Figure 10 examples can be seen of this same technique applied to images of microscopic breast cell tissue. For these images it is not as easy to tell whether the CNN is focusing on the right area of the image. In theory, malignant cell tissue is more complex and irregular, with nuclei and structures of varying sizes, compared to orderly and smooth benign tissue. However, without expert knowledge, verifying the CNN focus area is a difficult task. In an attempt to expand XAI methods for specifically breast cancer cell images, this project aims to research the viability of applying fractal analysis on CNN

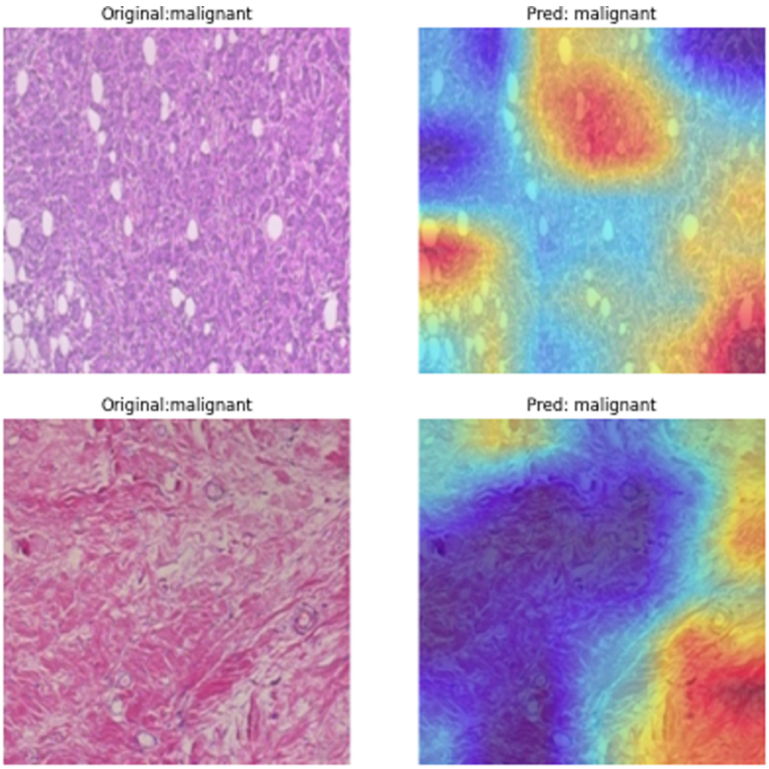


Figure 10: HiResCAM applied to breast cancer cell images (right) next to their original (left).

focus areas obtained by saliency maps to compute a value that is potentially easier to interpret for non-experts. This would allow for easier verification of CNN decision-making within this context and could potentially increase trust of non-experts in the models' classification. Not only that, improved interpretability of results could also contribute to a reduction of misclassifications, as well as play a big role in the education of new experts. The next section will discuss why fractal analysis could potentially fit the role for this measure.

3.2 The Role of Fractal Analysis

As stated previously, malignant cell tissue is on average more complex and irregular compared to benign cell tissue. In Section 2.4, it was concluded that fractal dimension can also be understood as a measure of roughness. This suggests that fractal dimension as an interpretable value, is suitable for the task at hand, as it theoretically should be able to discern malignant and benign cell tissue. This was indeed found in several studies. Chan and Tuszynski, for example, found that applying fractal analysis on 40x zoom histopathological images of breast cancer led to the promising result of a F1-score of 0.979 for benign vs malignant classification [CT16]. At higher magnification levels, the results were not significant in this study; nevertheless the findings seemingly support the use of fractal analysis within this context. Similarly, a study by Da Silva et al. [dSdSMdAM+21] showed that the computation of fractal dimension can be used to improve classification accuracy for the same 40x zoom breast cancer image dataset. The combination of CNNs with fractal dimension has also shown promise. In a study by Khan et al. [KWC+24] an ensemble of fractal analysis and

a CNN was created and applied to an early-onset Alzheimer detection task, about which they state: "The proposed fractional order-based CNN classifier achieves an improved accuracy of 99% as compared to the state-of-the-art models." Further support for the combination of CNNs and fractal analysis comes from a study by Roberto et al. [RLNZdN22], in which a CNN was separately trained on standard histopathology images, as well as feature maps containing fractal measures such as fractal dimension and lacunarity, leading to accuracies between 89.66% and 99.62%. This setup was applied to four different histology sets, including a breast cancer dataset, on which it reached the accuracy of 89.66%.

Based on these studies, fractal analysis and CNNs seem an appropriate and promising combination within the domain of breast cancer detection. Research into fractal analysis as an explainability tool, compared to performance improvement, is more limited but also shows potential. In the earlier referenced study by Khan et al. [KWC+24] their fractional order-based CNN is stated to have additional XAI capabilities, when compared to a standard CNN setup using the created feature maps. The angle of fractal analysis as a form of XAI was also explored in the study by Suraj [Sur25], who proposed a methodology for enriching CNNs with features, including fractal dimension, as an explainability tool. Their findings stated that the proposed pipeline did not affect classification and enhanced insight into the CNN decision-making process for a pneumonia detection task.

So in short, conducted research into fractal analysis as a method to classify breast cancer and other histology datasets has shown a lot of promise, both in combination with CNNs, as well as by itself. This project aims to expand on the research into the use of fractal dimension as an explainability tool for CNNs. Previous work under this angle has shown success for pneumonia-based tasks. In the next section, the methodology for testing fractal analysis as an explainability tool for CNN decision-making in breast cancer detection is described.

4 Methodology

In this section, the methodology for the experiment is described. Additionally, a breakdown is given of the design choices made and the options considered.

4.1 The Dataset

To train a CNN, a dataset is required that contains images with correct labels, so that the network can use these for its training. During this experiment, the BreakHis dataset was chosen. This dataset is publicly available and contains 9109 microscopic images of 82 patients. It shows breast cell tissue at different magnification levels with images being 700x460 pixels, 3-channel RGB, 8-bit depth in each channel, and in PNG format [SOPH16]. In Table 1, a breakdown of the distribution of the number of images per magnification level and label can be found. The images were obtained using the SOB method, also known as partial mastectomy. Besides distinguishing benign cell tissue and malignant cell tissue, the dataset also contains subtypes for both of these categories.

To fit the experiment, several changes to the dataset were made. For example, the choice was made to disregard the different subtypes for classification, and focus on the distinction between malignant

and benign structures. To be able to have generalizable results, as well as gain insight into the effect of the magnification factor, the choice was made to only use the images at 40x and 400x magnification. Choosing the magnification levels furthest apart allows for the best generalizability, without having to train four separate models. Additionally, the dataset was manually split into a train, validation, and test set. The train set is used with ground truth labels for the network to optimize its weights. The validation set is used during training, for the network to calculate its accuracy after each epoch with images not used for training. The test set is used after training is complete, to calculate the model's final accuracy. To achieve this split, the aim was to have approximately 75% of the images in the train set and 12.5% of the images in both the validation and test set. Even though subtypes were disregarded for classification, for balanced data splitting the images from the different subtypes were divided over the different splits following the same ratio where possible. Additionally, the constraint was set to have all images from a single patient in the same subset, preventing images that are very similar to each other from being spread between the training set and a set meant to evaluate performance without having seen the training images. In Tables 2 and 3, an in-depth breakdown of the split over the subsets can be found. In Figure 11 example images from the BreakHis dataset are shown.

Magnification	**Benign**	**Malignant**	**Total**
40X	652	1,370	1,995
100X	644	1,437	2,081
200X	623	1,390	2,013
400X	588	1,232	1,820
Total of Images	**2,480**	**5,429**	**7,909**

Table 1: Distribution of images by magnification in the BreakHis dataset.

Dataset/ Benign subtype	**Adenosis**	**Fibroadenoma**	**Phyllodes Tumor**	**Tubular Adenoma**	**Total**
Train	70	196	58	117	441
Test	29	35	13	16	93
Val	15	22	38	16	91
Dataset/Malignant subtype	**Ductal Carcinoma**	**Lobular Carcinoma**	**Mucinous Carcinoma**	**Papillary Carcinoma**	**Total**
Train	648	114	144	109	1015
Test	110	20	30	19	179
Val	106	22	31	17	176

Table 2: Overview of subtype division over train, test, and validation sets for benign (top) and malignant (bottom) images of 40x zoom.

After dataset retrieval and handling, the next step was image preprocessing, which could contribute to better CNN performance. During early testing, it was observed that the staining colors used during the SOB method resulted in differently shaded images, which could potentially become an unwanted feature used by the CNN to discern malignant and benign cell tissue. To prevent this, a

Dataset/ Benign subtype	Adenosis	Fibroadenoma	Phyllodes Tumor	Tubular Adenoma	Total
Train	60	194	59	100	413
Test	29	26	17	16	88
Val	17	17	39	14	87
Dataset/ Malignant subtype	**Ductal Carcinoma**	**Lobular Carcinoma**	**Mucinous Carcinoma**	**Papillary Carcinoma**	**Total**
Train	587	96	119	99	901
Test	102	24	26	15	167
Val	99	17	24	24	164

Table 3: Overview of subtype division over train, test, and validation sets for benign (top) and malignant (bottom) images of 400x zoom.

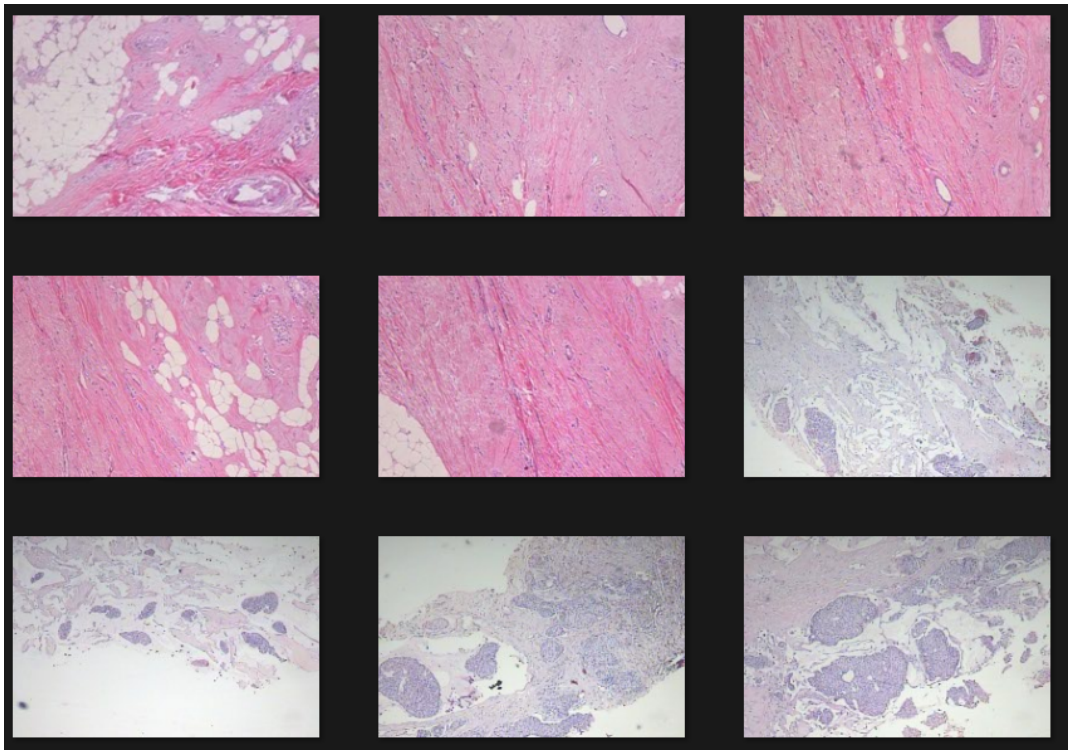


Figure 11: Example images from the BreakHIS dataset of breast cancer cell images.

color normalization method was applied, namely the Macenko method [MNM+09]. This method references an example image from the dataset and normalizes all images to the colors from the example. The result of this normalization process can be found in Figure 12.

After color normalization, data augmentation was applied. Data augmentation randomly applies selected feature changes to an image to prevent overreliance on specific features, as well as to increase the number of training images available. The data augmentations applied include resizing to 224x224 pixels (standard ResNet input size), random horizontal and vertical flips, and random rotations. Orientation changes were chosen as the orientation of the image should not affect the

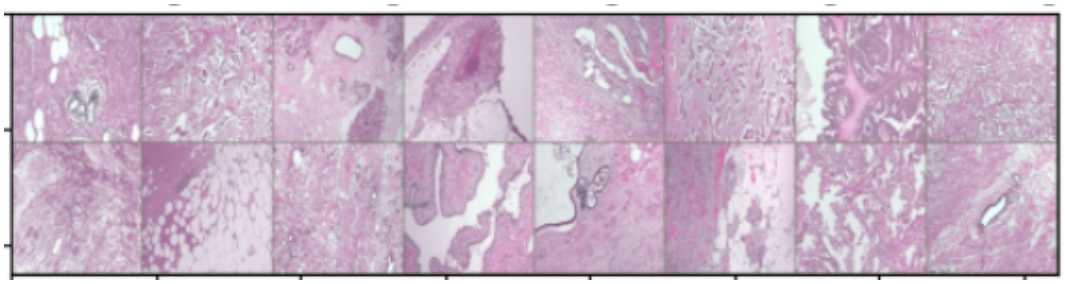


Figure 12: BreakHIS images after Macenko color normalization.

classification of breast cancer images. See Table 4 for an overview of the applied augmentations and their parameters.

Augmentation	**Parameter(s)**
Resize	224x224
RandomHorizontalFlip	p = 0.5
RandomVerticalFlip	p = 0.5
RandomRotation	15°
ToTensor	-
Normalize	Standard ImageNet mean and std

Table 4: Data augmentation overview.

4.2 The CNN

This section describes the chosen setup and the finetuning process of the CNN used for the classification of the BreakHis images during the experiment.

For the base of the network, the approach of transfer learning was chosen. Transfer learning involves importing a pre-trained CNN with its weights and structure and finetuning it on the chosen dataset, so that it can adapt to the task it is needed for. For this project in particular, a ResNet-50 model was chosen pretrained on the ImageNet dataset. ResNet-50 is a 50-layer-deep convolutional neural network, which was introduced in 2015 [HZRS16] with as main improvement over other CNNs the use of residual blocks. Residual blocks are a solution to the vanishing gradient problem by using skip connections for improved gradient flow. This way the model can make more efficient use of its deeper layers. ResNet-50 is a common choice for transfer learning tasks as it has been shown to be both efficient and accurate, making it very suitable as a backbone [AYM22]. During the experimentation phase, the VGG-16 model was tested as well, but did not show an improvement over the chosen setup.

ImageNet is a dataset of more than 14 million images of objects and scenes from everyday life [DDS+09]. Even though, ImageNet images and breast cancer classification images are not similar, the extensive training allows the model to be good at structural tasks such as edge detection and segmentation from the start. As a result, training time is reduced significantly and only finetuning for the classification of benign and malignant breast cancer cells is required.

The difference in datasets between BreakHIS and ImageNet was the reason to make all CNN layers trainable. The alternative of freezing most layers and retraining just the classifier or a few layers is better suited for more similar databases. Experimentation also showed that unfreezing all layers improved performance significantly, making this the preferred method. Additionally, dropout was applied to the model. This means that at each training step a percentage of neurons (set to 33%) is switched off to prevent overfitting on the training data.

For the loss function, Binary Cross-Entropy with Logits (BCEWithLogitsLoss) was chosen. The loss function is the designated way for the model to calculate its performance. Since this is a binary task, BCEWithLogitsLoss is the standard approach, whereas CrossEntropyLoss is the standard for multiclass classification. The optimizer is the algorithm that is used for updating the weights of the model, of which several options exist. Based on experimentation with the Adam, AdamW and SGD optimizers and the recommendation from Agarwal et al. [AYM22], the Adam optimizer was chosen for the CNN. Finally, a scheduler was chosen. The scheduler controls the learning rate of the model during training. Having this at the right values can help speed up training, avoid local minima and reduce overfitting. During experimentation OneCycleLR, CosineAnnealingLR and ReduceLROnPlateau were tried, of which CosineAnnealingLR performed best. This scheduler adjusts the learning rate according to a cosine function, leading to different stages of speeding up training and utilizing learned features. In Table 5 an overview is given of the chosen options and the used parameters.

This concludes the section on the selection of options and parameter fine-tuning for the CNN used to classify the BreakHis images, and thus the first step of the methodology.

Component	**Choice**	**Parameters**
Criterion	BCEWithLogitsLoss	pos_weight: class normalization
Optimizer	Adam	learning rate = 1e-5 weight_decay = 1e-4
Scheduler	CosineAnnealingLR	T_max = 30 eta_min = 1e-6

Table 5: Training configuration: criterion, optimizer, and LR scheduler.

4.3 Saliency Maps

As discussed previously, Grad-CAM is one of the most common methods to create a saliency map and was therefore considered for the extraction of the focus area of the CNN. With the use of the PyTorch Grad-CAM package access to several more saliency map methods was available as well. Out of these, the following were considered: Grad-CAM++, which uses second-order gradients in its calculation instead of first-order; HiResCAM, which improves on Grad-CAM by using element-wise calculation instead of averaging; Score-CAM, which applies perturbations and constructs maps based on confidence scores; and LayerCAM, which weighs activations by positive gradients [GHY+23]. Based on its simplicity, improvement over standard Grad-CAM, and best performance in early testing, the decision was made to use HiResCAM for extracting the CNN focus area. The saliency maps previously shown in Figure 10 were made using this method.

4.4 Masking

After obtaining the saliency maps, initially the idea was to take the area of highest CNN attention according to the heatmap and isolate it, as shown in Figure 13a. Then, the obtained area was inverted to create the non-focus mask. However, experimentation showed that due to the way the box-counting method, described in Section 2.4, works, the size and shape of the created masks heavily influence the computed fractal dimension. As a result, comparing focus and non-focus areas was strongly influenced by the larger size and substantially different shape of the non-focus areas.

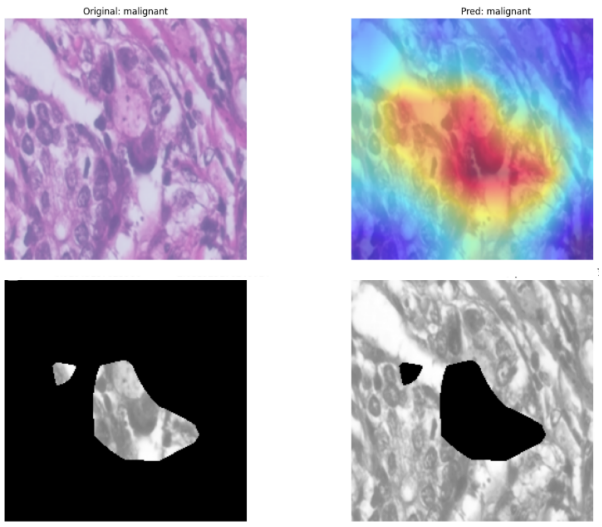
To prevent this from happening the following two methods were tried: the first idea was to use the same focus area mask, but instead of inverting for the non-focus area mask, a copy of the focus area mask was made and shifted on top of a suitable location in the non-focus area. This way, the masks for the focus and non-focus area of the same image are of the same size and shape, allowing for accurate comparison. See Figure 13b for an example. When comparing benign vs malignant images however, this still causes issues as the focus mask shapes still vary considerably between images. The second idea was therefore to have a standard sized box that is placed around the pixel with the highest attention level in the saliency map and use that as focus area. For the non-focus area the same-sized box can be put on the lowest activation area of the heatmap, keeping the masks consistent (see Figure 13c). Even though this introduces new challenges such as possible overlap and determining an appropriate box size, the choice was made to use this version of the masks during the final experiment, as it would allow for the fairest comparison.

4.5 Pattern analysis

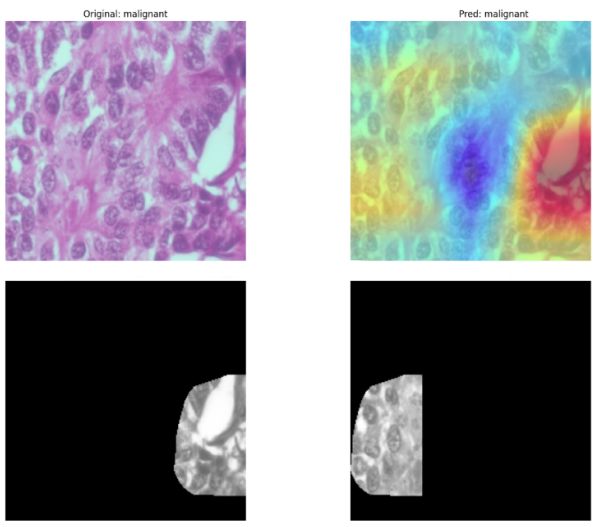
The next step was to apply fractal analysis and other pattern analysis methods to the computed focus and non-focus areas. To determine fractal dimension, a modified version of the box-counting method from Section 2.4 was used. Usually, box-counting counts every box that has any part of the pattern of interest in it. However, for histopathological images there is structure and matter distributed over the entire image, making binary counting uninformative. At first, skeletonization of the image was attempted, however isolating the correct cell tissue using this method was deemed too difficult. Eventually, the choice was made to use a differential box-counting method. This method estimates fractal dimension using the intensity variation between scales, where intensity is defined as the difference in pixel value between the highest and lowest value in the box [SC94].

Besides fractal dimension, lacunarity was computed as well. The applied function computes lacunarity in similar boxes as the fractal dimension function, using the following formula: $\Lambda = \frac{\mathrm{Var}(M)}{(\mathbb{E}[M])^2}$, where M is the mass in the box, Var the variance and $\mathbb{E}$ the mean. The average over the image is returned as the lacunarity.

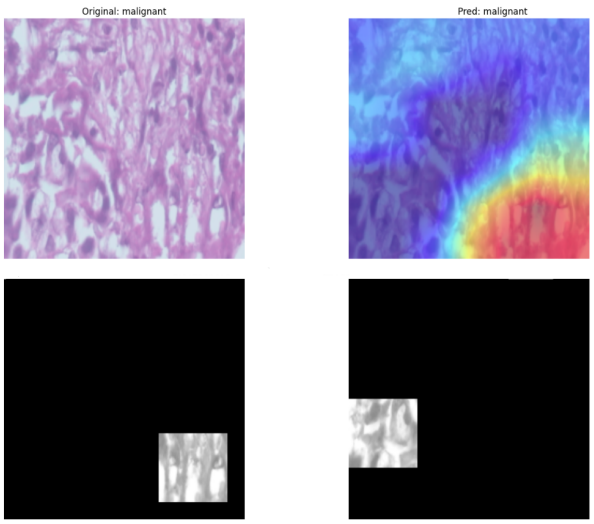
Two more measures were computed for the sake of comparison. Local binary patterns (LBP) and gray-level co-occurrence matrices (GLCM) are two commonly used pattern analysis tools for histopathological images [PS23, ÖA18]. LBP compares pixel values to their neighbors and creates a histogram showing the spread [SYCZ13]. Using the Shannon Entropy of the histogram, which is a measure of randomness, the complexity of the LBP values was calculated and used for further analysis. GLCM measures the joint occurrence of intensity pairs and puts it in a matrix [SSA17]. Based on this matrix, different features can be used for further analysis. For this project the decision



(a) Focus: CAM area mask,
Non-focus: CAM area inverted.



(b) Focus: CAM area mask,
Non-focus: CAM area shifted.



(c) Focus: Bounding box highest CAM value,
Non-focus: Bounding box lowest CAM value.

Figure 13: Overview of the considered masking options, with focus masks on the left and non-focus masks on the right.

was made to use contrast and homogeneity, as indicators of complexity.

4.6 Statistical Tests

With the focus and non-focus areas determined and the measures of interest computed, the final step was to apply statistical tests for comparison of focus and non-focus areas, as well as benign and malignant images. The first goal was to test whether for each measure there is a significant difference between the values computed in the focus area and those computed in the non-focus

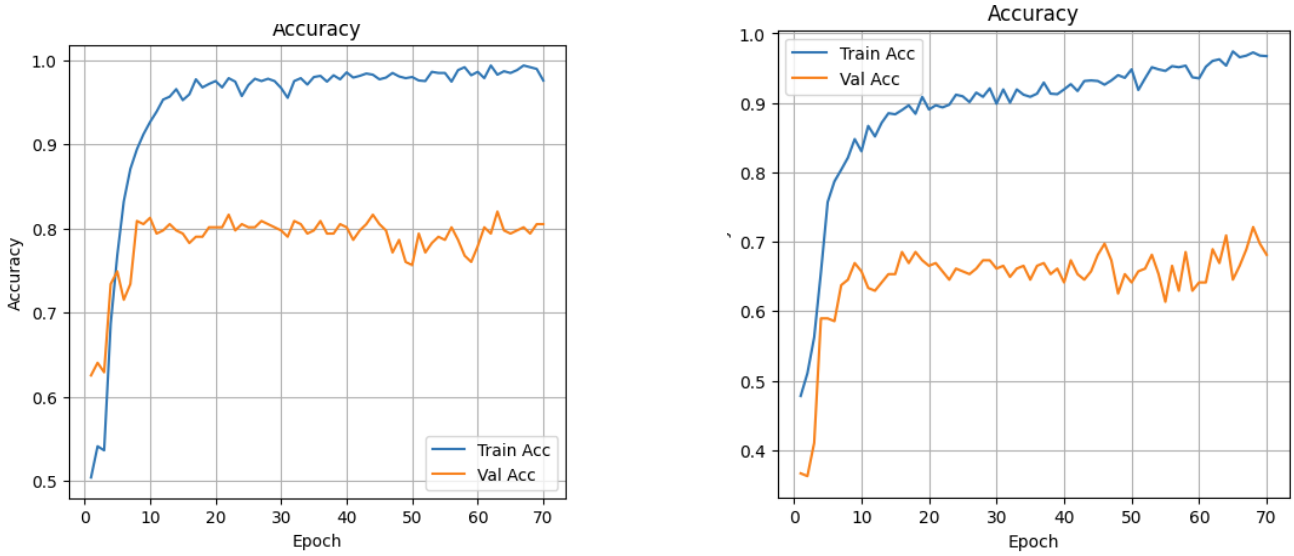
area. Since a focus area value and a non-focus area value are computed from each image, this is a paired analysis. As a result, the Wilcoxon signed-rank test was used [KE19]. It is expected that the test will find significant differences between focus and non-focus areas, as the CNN should focus on malignant cell tissue, which is in general rougher and more complex compared to benign cell tissue. For the comparison of the values obtained in malignant images to the values in benign images, it is expected that based on values extracted from focus areas there is a significant difference for the same reason as before. However, when making the same comparison for non-focus areas it is expected that this is not necessarily the case. Comparing benign to malignant images is an independent task, as the values are computed from different images. For that reason, the Mann-Whitney U test was used for this task [KE19].

This section has described the methodology used for the experiment, as well as discussed relevant design choices. In the next section, the results and their interpretation can be found.

5 Results and Interpretation

5.1 Results

The full CNN performance can be found in Table 6. For the 40x zoom images on the test dataset, that had not interacted with the model during training, a final accuracy of 83.46% was found. Additionally, the precision was calculated to be 83.84% and the recall 92.74%. For the 400x zoom dataset, the test set accuracy was 80.00% with precision at 85.80% and recall at 83.23%. In Figure 14 an overview of the computed accuracy throughout training can be found for both these datasets.



(a) 40× magnification

(b) 400× magnification

Figure 14: Training and validation accuracy per epoch for CNN models trained on BreakHIS images at different magnification levels.

Magnification	Dataset	Accuracy (%)	Precision (%)	Recall (%)
40×	Train	99.93	100.00	99.90
	Validation	82.02	84.41	89.20
	Test	83.46	83.84	92.74
400×	Train	99.16	100.00	98.78
	Validation	72.11	80.52	75.61
	Test	80.00	85.80	83.23

Table 6: Model performance across training, validation, and test sets for 40× and 400× magnification.

The main results of the experiment can be found in Table 7. It shows the obtained p-values of comparing the computed fractal dimension in focus areas vs non-focus areas and malignant vs benign images using the Wilcoxon signed-rank test for the former and Mann-Whitney U test for the latter. The shown results were computed on the focus areas obtained using the square mask method. Additionally, results of computation on both the test set and the validation set are shown for both magnification level datasets. The former will be the main focus in the next section in which an interpretation is given, because it contains the images not used during training of the CNN. The latter is included to give an insight into the level of consistency of the pipeline.

In Tables 8 and 9 found in the Appendix, a full overview of all done experiments can be found. They show computed p-values not only for fractal dimension, but also for lacunarity, local binary patterns (LBP) and the gray-level co-occurrence matrix (GLCM) on both the test and validation sets, which allows for comparison to industry-standard methods. Additionally, these tables show the results for correctly classified images for further insight.

Magnification	Comparison	Dataset	P-value
40×	Focus vs Non-Focus (Wilcoxon)	Malignant Test	0.0233
		Malignant Val	2.20e-4
		Benign Test	0.998
		Benign Val	0.366
40×	Benign vs Malignant (Mann-Whitney U)	Focus areas Test	0.0102
		Focus areas Val	0.0635
		Non-focus areas Test	0.0959
		Non-focus areas Val	0.336
400×	Focus vs Non-Focus (Wilcoxon)	Malignant Test	5.69e-08
		Malignant Val	2.63e-08
		Benign Test	2.45e-05
		Benign Val	7.85e-05
400×	Benign vs Malignant (Mann-Whitney U)	Focus areas Test	0.823
		Focus areas Val	0.435
		Non-focus areas Test	0.227
		Non-focus areas Val	0.0803

Table 7: P-values from Wilcoxon signed-rank tests (Focus vs Non-Focus) and Mann–Whitney U tests (Benign vs Malignant) for 40x and 400x magnification datasets.

5.2 Interpretation

In this section, an interpretation is given for the results shown previously. Figure 14 shows training and validation accuracy of both datasets per epoch. In both graphs, it can be seen that validation accuracy stabilizes in just under 10 epochs, while training accuracy continues to increase in subsequent epochs, reaching near 100%. The computed test set accuracies as found in Table 6 are slightly lower than more complex set-ups from performance-focused related work [AYM22, EBI24, SBW24], but show a similar trend of the 40x dataset performing better than the 400x dataset. See Section 6 for further discussion on possible CNN improvement, as well as the signs of overfitting. Nevertheless, the found performance seems sufficient for the rest of the pipeline.

The main results of interest to answer the research question can be found in Table 7. The shown p-values will be evaluated at a significance level of 0.05. For the 40x dataset, the Wilcoxon signed-rank test that compared focus areas to non-focus areas on malignant images gave a p-value of 0.0233, which is less than 0.05, indicating that the computed fractal dimension values differ significantly between the compared groups. These results suggest that the CNN may focus on areas of the image with higher complexity, and that this can be represented by fractal dimension for images containing malignant cell tissue. These findings support the idea of using fractal dimension as an explainability tool for CNNs on breast-cancer detection tasks. As a control test, the same Wilcoxon signed-rank test, comparing focus and non-focus areas, was applied to benign images as well. Here, the p-value was significantly higher than 0.05, indicating no significant difference between focus and non-focus areas for benign images, according to expectation.

For the independent Mann-Whitney U test comparing benign images vs malignant images using the focus areas, a p-value lower than 0.05 was found for the test set. For the validation set the p-value was calculated to be slightly above the desired significance level. As a result, these findings should be interpreted with caution. Nevertheless, the results suggest a significant difference between the focus areas of benign images and those of malignant images, as was expected. The control test comparing non-focus areas between benign and malignant images yielded p-values larger than the significance level, supporting the hypothesis that complexity, and thus fractal dimension, is similar between these two groups.

In the second half of Table 7, the results for the 400x dataset can be found. It shows that every focus vs non-focus comparison showed significant differences, while every benign vs malignant comparison showed no significant difference at the chosen significance level. As a result, it is not possible to conclude that fractal dimension is a reliable explainability tool for the CNN output on the 400x dataset. A similar conclusion was reached in studies by Roberto et al. and Chan and Tuszynski, who also found significant results for using fractal dimension and other fractal measures on the BreakHis 40x magnification dataset, but not on the 400x dataset [RLNZdN22, CT16].

The same methodology was applied using other pattern analysis methods instead of fractal dimension, including lacunarity, LBP and GLCM. The results for these measures, as well as some other computed metrics can be found in Tables 8 and 9 in the appendix. These results indicate that fractal dimension is the only measure that largely shows significant differences where expected and non-significant differences in the control tests. Additional discussion on the performance of the alternative pattern analysis methods will follow in the next section. Nevertheless, the presented results point towards fractal dimension being capable of distinguishing malignancies from benign tissue and therefore

being a suitable measure for increasing insight into CNN output, when combined with the saliency map method, for 40x zoom images specifically. Further discussion on design choices, sensitivity of results and points of potential improvement will follow in the next section.

6 Discussion

During the experiment, several methods were tried and design choices made. This section gives an overview of choices that worked out well and areas for potential improvement.

To start off, the BreakHis dataset is a commonly used open access dataset of JPG format breast cancer images. This dataset gives a good starting point for training a CNN on the task of breast cancer classification and formed a suitable dataset for the performed research. However, for further research it would be recommended to use different datasets to test the generalizability of the results. In particular, the use of whole slide images (WSI) would potentially allow for new insights. WSI is a high-resolution, multiscale image representation, potentially making pattern analysis more effective at discriminating benign and malignant tissue.

The next dataset-related point of discussion is the test set achieving better performance compared to the validation set after training. The validation set is used during training for hyperparameter tuning, while the test set is used to check final accuracy at the end. As a result, the observed higher accuracy, especially for the 400x magnification dataset, on the test set as seen in Table 6 was not expected. It is hypothesized that this result is possibly caused by the benign and malignant subtypes available in the original dataset. During preprocessing, an attempt was made to split every provided cell subtype over the train, validation, and test sets according to the global 75/12.5/12.5 split. However, images from a single patient are not allowed to be split over different sets. This is to prevent less accurate evaluation caused by training and evaluation images being too similar. However, as a result of this constraint, subtypes with available images from only a few patients were impossible to split according to the global split. Therefore, it might be possible that subtypes were underrepresented in the test or validation set, resulting in a comparatively more difficult classification task. Refer back to Tables 2 and 3 for the currently applied split. For future research, a larger dataset or extra augmentation could be used to potentially prevent the present unbalance.

For the CNN performance itself, a final accuracy of 83.46% was reached on the 40x dataset. The current performance was accomplished using a ResNet-50 architecture with ImageNet pre-training. In studies by Agarwal et al. and Simonyan et al. [AYM22, SBW24] it is suggested that a DenseNet or VGG architecture could potentially achieve higher accuracy. During the experimentation phase, a VGG-16 model was briefly tested, but it did not perform better than the current model. In theory, a ResNet-50 model is a great fit, achieving high accuracy at a relatively low computational cost using transfer learning, making it the model of choice. Nevertheless, for further research, different models and/or optimization of the CNN performance could be an avenue to explore. For example, in the mentioned study by Agarwal et al., a final accuracy between 88% and 95% was found depending on the model [AYM22]. Other studies with more complex CNN setups also reached similar performance [EBI24, SBW24, RLNZdN22]. A point of focus to potentially bridge this gap, is the presence of signs of overfitting, for example the discrepancy between training and validation accuracy in Figure 14 and the increasing validation loss. Dropout, batch normalization, weight

decay, data augmentations, and a learning rate scheduler have already been applied within the pipeline. However, additional augmentation, gradual unfreezing, learning rate optimization, and early stopping could be explored in further research to improve performance. The achieved accuracy was deemed sufficient for further experimentation. Nevertheless, an improved CNN setup could lead to new insights and a more robust model.

For the extraction of focus areas from the CNN, the methodology relies on a saliency map method, namely HiResCAM. However, there exists a discussion about the reliability of saliency map methods for CNNs in general. Based on the results of the experiment and it being a commonly applied XAI method, this appeared to be a appropriate choice. Nevertheless, it has to be acknowledged that for example Graziani et al. concluded that: "The qualitative evaluation alone is thus not sufficient to establish the appropriateness and reliability of the visualization tools" [GLMA20], based on a study on common XAI methods for histopathology, including Grad-CAM. A supporting study in favor of this technique is for example Dörrich et al., who found that CAM-maps generated by Grad-CAM produce focus areas that generally coincide with expert judgment, based on an experiment focused on head and neck cancer histopathology images [DHF+23]. As a result, HiResCAM as an improved version of Grad-CAM [DC20] was deemed a good fit for the task. Nevertheless, as better XAI methods become common ground, they could improve the proposed pipeline from this project.

Another important design point within the pipeline is the masking of the focus and non-focus areas as described in Section 4.4. As discussed there, taking the most intuitive approach of using the high focus area of the heatmap as focus area and the inverse as the non-focus area led to misleading results. This is because fractal dimension is influenced by the shape and size of the mask. To control for this a version was tested where the focus area was the same, but the non-focus area consisted of the focus area mask shifted on top of a part of the non-focus area. This keeps size and shape of the tested area within each image consistent, resulting in an equal comparison of focus vs non-focus area. However, between images the size and shape of the masks still differed largely, making comparison between benign and malignant images difficult. The final solution to use a box of a set size around the area of highest importance proved the most effective and achieved the reported results. This method, however, requires choosing a box size, which can affect results, and introduces the possibility of overlap between focus and non-focus areas. The latter is unlikely, as the boxes are set around the highest and lowest pixel values respectively, and it did not occur during the experiment, but this possibility must be acknowledged. Due to time constraints, a box size of similar size to the average focus-area-shaped mask was chosen. For future research, it is recommended to explicitly calculate the average CAM-area-mask size or explore other masking methods that were not considered in this study.

For the pattern analysis part of the pipeline, skeletonizing the image was considered. This is an operation often combined with the computation of fractal dimension as it reduces structures to single-pixel lines to leave a clear structure to analyze. However, during experimentation it showed that getting the right skeleton was very difficult for the breast cancer images. It would lose out on vital information by filtering out important structure from the images. The differential box-counting method as described in Section 2.4 was used instead to capture the structure of interest in a different manner. However, this method can also suffer from inaccuracies, such as over- or under-counting boxes and inefficiencies, according to a review by Panigrahy et al. [PSMB19]. As a result, evaluating alternative box-counting algorithms could be a valuable angle for future research.

As can be seen from the above discussion, a lot of factors can influence results and have to be right for the pipeline to work. Most of the time and effort of this research have been focused on fractal dimension in particular as an analysis tool. Applying the same methodology to the other considered pattern analysis methods of lacunarity, LBP and GLCM did not achieve meaningful results (see Tables 8 and 9). This could be interpreted as support for using fractal dimension for further applications. However, it is also possible that the current pipeline is not optimal for the alternative analysis methods. During experimentation, several iterations of each of these methods were tried and an attempt was made to have them work, nevertheless the main focus has been on fractal dimension from the start, which could be an alternative explanation for these methods not being successful.

For the statistical analysis part of the pipeline, initially Cohen's D was computed in an attempt to capture effect size of the results. As a consequence, this measure is still visible within the code and in Tables 8 and 9. However, late in the research I realized that Cohen's D is not suited for the used Wilcoxon signed-rank test and Mann-Whitney U test, as these are based on median change, whereas Cohen's D is meant for the comparison of means. As a result, a strong recommendation would be to compute the rank-biserial correlation coefficient as effect size instead, which was not done as a result of time constraints.

Finally, and most importantly, a discussion on the sensitivity of the results will follow. Throughout the process, many options and possibilities were tried with differing success. However, it was noticed that small changes could affect the final results substantially. A clear example of this is the different masking possibilities as explained, but also things such as applied augmentations, mask size and retraining the CNN. As a result, the decision was made to report the first output of the completed pipeline, as found in Section 5, to provide an objective overview. However, based on this observation, the recommendation would be to further explore the viability of fractal dimension as an explainability tool for CNN decision-making for breast cancer images. A more robust and refined pipeline, building upon the conducted experiment, could lead to a more definitive conclusion on the viability of fractal dimension as an explainability tool. Nevertheless, the belief is that the presented literature, as well as the obtained results, support the idea of using fractal dimension in this manner. If this would indeed be the case, current explainability methods could be expanded with a measure currently underexplored within the histopathological field, potentially contributing to breaking down the black-box problem of the extremely promising CNN technique.

7 Conclusion

So to conclude, in the pursuit of improving the treatment of one of the most common forms of cancer, breast cancer, convolutional neural networks can and are already starting to play a big role. Based on large amounts of data, these models are trained to learn and classify malignant and benign cell tissue with very high accuracies, creating a very useful tool for doctors to speed up and improve their decision-making. However, for the use of systems like these, especially in contexts as important as the medical field, understanding why a certain decision was made is also very important. The standard CNN decision-making often lacks this insight, operating in a black-box fashion. Common explainability methods for CNNs are also not particularly effective when applied within this field. For example, saliency maps require expert knowledge for the interpretability of

the method output.

In this research, a CNN was trained on the open-source breast cancer dataset BreakHis and a saliency map was applied to its classifications to extract the CNN focus areas. From the obtained focus and non-focus areas, fractal dimension and several other measures were computed and compared between focus and non-focus areas and between benign and malignant images. Using this pipeline, the goal was to answer the question whether fractal analysis, and in particular fractal dimension, is suitable as a computable and understandable measure to increase insight into CNN decision-making for breast cancer histopathological images. It was found that, for 40x zoom images, there were significant differences in values between the focus areas of malignant images compared to those of benign images and between focus and non-focus areas. Since fractal dimension can intuitively be described as a measure of roughness or complexity of a pattern, these results support the notion of fractal dimension potentially being useful as an easy-to-understand measure to increase insight into CNN decision-making on breast cancer cell classification tasks. An important observation is that the potential improvements and the sensitivity of the results to the numerous possible design choices indicate a need for further research. Nevertheless, in an effort to use technology to save lives, fractal dimension as a measure of complexity could possibly contribute to making the technology more interpretable and trustworthy for everyone involved.

References

[ABV+20] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20, 2020.

[AYM22] P. Agarwal, A. Yadav, and P. Mathur. Breast cancer prediction on BreakHis dataset using deep CNN and transfer learning model. In P. Nanda, V. K. Verma, S. Srivastava, R. K. Gupta, and A. P. Mazumdar, editors, *Data Engineering for Smart Systems*, volume 238 of *Lecture Notes in Networks and Systems*. Springer, Singapore, 2022.

[CT16] A. Chan and J. A. Tuszynski. Automatic prediction of tumour malignancy in breast cancer with fractal dimension. *Royal Society Open Science*, 3, 2016.

[Dal00] M. Dale. Lacunarity analysis of spatial pattern: A comparison. *Landscape Ecology*, 15:467–478, July 2000.

[DC20] R. L. Draelos and L. Carin. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020.

[DDS+09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[DHF+23] M. Dörrich, M. Hecht, R. Fietkau, et al. Explainable convolutional neural networks for assessing head and neck cancer histopathology. *Diagnostic Pathology*, 18:121, 2023.

[dSdSMdAM+21] L. G. da Silva, W. R. S. da Silva Monteiro, T. M. de Aguiar Moreira, et al. Fractal dimension analysis as an easy computational approach to improve breast cancer histopathological diagnosis. *Applied Microscopy*, 51:6, 2021.

[EBI24] R. B. Eshun, M. Bikdash, and A. K. M. K. Islam. A deep convolutional neural network for the classification of imbalanced breast cancer dataset. *Healthcare Analytics*, 5:100330, 2024.

[FPDS99] K. Foroutan-Pour, P. Dutilleul, and D. L. Smith. Advances in the implementation of the box-counting method of fractal dimension estimation. *Applied Mathematics and Computation*, 105(2–3):195–210, 1999.

[GHY+23] X. Guo, B. Hou, C. Yang, S. Ma, B. Ren, S. Wang, and L. Jiao. Visual explanations with detailed spatial information for remote sensing image classification via channel saliency. *International Journal of Applied Earth Observation and Geoinformation*, 118:103244, 2023.

[GLMA20] M. Graziani, T. Lompech, H. Müller, and V. Andrearczyk. Evaluation and comparison of CNN visual explanations for histopathology. In *Proceedings of the XAI Workshop at AAAI-21*, 2020.

[HC24] J. Halamka and P. Cerrato. Using AI to predict the onset of sepsis, May 2024.

[HZRS16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[KE19] A. P. King and R. J. Eckersley. Inferential statistics iii: Nonparametric hypothesis testing. In A. P. King and R. J. Eckersley, editors, *Statistics for Biomedical Engineers and Scientists*, pages 119–145. Academic Press, 2019.

[KRM+24] S. Khalighi, K. Reddy, A. Midya, K. B. Pandav, A. Madabhushi, and M. Abedalthagafi. Artificial intelligence in neuro-oncology: Advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. *NPJ Precision Oncology*, 8:80, 2024.

[KUA+25] G. Komala, S. Umar, P. C. R. Alla, K. Vaghela, N. T. Gole, and N. P. Bodne. Competitive analysis for CNN fusion methodology with fractal feature extraction for image classification of brain health. In V. Bhateja, V. Abdul Hameed, S. K. Udgata, and J. Tang, editors, *Innovations in ICT: Sustainability for Societal and Industrial Impact*, pages 249–259. Springer Nature Singapore, 2025.

[KWC+24] Z. A. Khan, M. Waqar, N. I. Chaudhary, M. J. A. A. Raja, S. Khan, F. A. Khan, I. I. Chaudhary, and M. A. Z. Raja. Fractional gradient optimized explainable convolutional neural network for Alzheimer's disease diagnosis. *Heliyon*, 10:e39037, 2024.

[Man83] B. B. Mandelbrot. *The Fractal Geometry of Nature*. W.H. Freeman, San Francisco, 1983.

[Man06] B. B. Mandelbrot. Fractal analysis and synthesis of fracture surface roughness and related forms of complexity and disorder. *International Journal of Fracture*, 138:13–17, 2006.

[MCAS+22] H. Moujahid, B. Cherradi, M. Al-Sarem, L. Bahatti, A. B. A. B. M. Y. Eljialy, A. Alsaeedi, and F. Saeed. Combining CNN and Grad-CAM for COVID-19 disease prediction and visual explanation. *Intelligent Automation & Soft Computing*, 2022.

[MNM+09] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1107–1110, 2009.

[NJS+18] K. Nagasubramanian, S. Jones, A. K. Singh, A. Singh, B. Ganapathysubramanian, and S. Sarkar. Explaining hyperspectral imaging based plant disease identification: 3D CNN and saliency maps. *arXiv preprint arXiv:1804.08831*, 2018.

[NRN25] F. Nasir, S. Rahman, and N. Nasir. Breast cancer detection using convolutional neural networks: A deep learning-based approach. *Cureus*, 17(5):e83421, 2025.

[ÖA18] Ş. Öztürk and B. Akdemir. Application of feature extraction and classification methods for histopathological image using GLCM, LBP, LBGLCM, GLRLM and SFTA. *Procedia Computer Science*, 132:40–46, 2018.

[PS23] K. Prakash and S. Saradha. Efficient prediction and classification for cirrhosis disease using LBP, GLCM and SVM from MRI images. *Materials Today: Proceedings*, 81(Part 2):383–388, 2023.

[PSMB19] C. Panigrahy, A. Seal, N. K. Mahato, and D. Bhattacharjee. Differential box counting methods for estimating fractal dimension of gray-scale images: A survey. *Chaos, Solitons & Fractals*, 126:178–202, 2019.

[RLNZdN22] G. F. Roberto, A. Lumini, L. A. Neves, and M. Zanchetta do Nascimento. Fractal neural network: A new ensemble of fractal geometry and convolutional neural networks for the classification of histology images. *Expert Systems with Applications*, 166, 2022.

[SBW24] E. O. Simonyan, J. A. Badejo, and J. S. Weijin. Histopathological breast cancer classification using CNN. *Materials Today: Proceedings*, 105:268–275, 2024.

[SC94] N. Sarkar and B. B. Chaudhuri. An efficient differential box-counting approach to compute fractal dimension of image. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(1):115–120, Jan. 1994.

[SCD+17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[SOPH16] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016.

[SSA17] S. Singh, D. Srivastava, and S. Agarwal. GLCM and its application in pattern recognition. In *2017 5th International Symposium on Computational and Business Intelligence (ISCBI)*, pages 20–25, 2017.

[Sur25] T. Suraj. Interpretable artificial intelligence with explainability and robustness in medical image classification using topological and fractal features. *International Journal of Artificial Intelligence & Machine Learning*, 4:43–68, 2025.

[SVZ14] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

[SYCZ13] K.-C. Song, Y.-H. Yan, W.-H. Chen, and X. Zhang. Research and perspective on local binary pattern. *Acta Automatica Sinica*, 39(6):730–744, 2013.

[VTA+19] D. Varshni, K. Thakral, L. Agarwal, R. Nijhawan, and A. Mittal. Pneumonia detection using CNN based feature extraction. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–7, 2019.

[Wor24] World Cancer Research Fund. Worldwide cancer data. https://www.wcrf.org/preventing-cancer/cancer-statistics/worldwide-cancer-data/, 2024. Accessed: 14-12-2025.

[ZRSC15] Á. Zarándy, C. Rekeczky, P. Szolgay, and L. O. Chua. Overview of CNN research: 25 years history and the current trends. In *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 401–404, 2015.

A Code and Acknowledgments

Find the used code down below. 40x Pipeline: https://colab.research.google.com/drive/1JP8kDvECivfAWX3_PnG6ZJKtais1XMOl?usp=sharing 400x Pipeline: https://colab.research.google.com/drive/1FF1R76ckcpO79QBJJuMrioEKuY8_SIyz?usp=sharing

Due to size, the modified BreakHIS dataset is not provided. The full BreakHIS dataset can be retrieved from: https://www.kaggle.com/datasets/ambarish/breakhis

The base of the CNN structure was retrieved from: https://github.com/rabby0101/ResNet-50

The Macenko color normalization code was retrieved from: https://www.geeksforgeeks.org/machine-learning/macenko-method-for-normalizing-histology-slides-for-quantitative-analysis

The HiResCAM functionality was retrieved from the Grad-CAM package on: https://github.com/edw008/pytorch-grad-cam

The fractal dimension and lacunarity code was based on: https://xbe.at/index.php?filename=Fractal+Dimension+Analysis+of+Images+using+Python.md

ChatGPT was used during the experimentation phase as an assistance tool for code, such as helping with debugging of the code that was combined and adapted by me from the above sources. During the writing of the report ChatGPT was solely used to check grammar, spelling, punctuation and provide assistance with things such as table layout and LaTeXformatting.

B Full Results Tables

Feature	P-value (test-set)	Cohen's D (test-set)	P-value (val-set)	Cohen's D (val-set)
Focus vs Non-focus (Malignant)				
Fractal Dimension	0.0233	0.172	0.00022	0.274
Lacunarity	3.92e-28	-1.48	1.66e-26	-1.34
Lbp	0.0613	-0.125	0.0121	-0.157
Glcm_contrast	0.593	-0.1	0.00589	0.198
Glcm_homogeneity	0.787	-0.044	0.0325	-0.164
Focus vs Non-focus (Benign)				
Fractal Dimension	0.998	0.0147	0.366	0.0134
Lacunarity	1.27e-09	-1.13	5.94e-10	-1.04
Lbp	0.337	-0.113	0.0785	0.245
Glcm_contrast	0.0687	-0.201	0.792	0.0475
Glcm_homogeneity	0.0824	0.237	0.211	0.223
Focus vs Non-focus (Correct Malignant)				
Fractal Dimension	0.0227	0.179	0.000368	0.271
Lacunarity	6.22e-26	-1.45	3.52e-25	-1.37
Lbp	0.028	-0.154	0.012	-0.163
Glcm_contrast	0.437	-0.123	0.00206	0.229
Glcm_homogeneity	0.71	-0.0378	0.00772	-0.182
Focus vs Non-focus (Correct Benign)				
Fractal Dimension	0.78	0.0144	0.32	0.055
Lacunarity	5.42e-07	-0.964	1.25e-05	-0.818
Lbp	0.586	-0.077	0.715	0.0947
Glcm_contrast	1.26e-05	-0.238	2.54e-06	-0.0352
Glcm_homogeneity	0.041	0.335	0.157	0.281
Benign vs Malignant (Focus)				
Fractal Dimension	0.0102	-0.382	0.0635	-0.317
Lacunarity	0.00831	0.766	1.48e-06	0.816
Lbp	0.000519	-0.531	0.728	0.064
Glcm_contrast	0.00069	-0.431	1.26e-05	-0.593
Glcm_homogeneity	0.00031	0.549	5.77e-05	0.481
Benign vs Malignant (Non-Focus)				
Fractal Dimension	0.0959	-0.214	0.336	0.0363
Lacunarity	0.039	0.496	0.00143	0.587
Lbp	0.000779	-0.517	0.00238	-0.392
Glcm_contrast	0.0147	-0.267	0.00412	-0.383
Glcm_homogeneity	0.0136	0.225	0.0628	0.0413
Benign vs Malignant (Correct Focus)				
Fractal Dimension	0.0121	-0.464	0.0434	-0.341
Lacunarity	7.99e-05	1.18	1.42e-07	1.24
Lbp	0.059	-0.374	0.949	-0.00362
Glcm_contrast	2.67e-05	-0.61	8.08e-06	-0.74
Glcm_homogeneity	2.16e-06	0.802	1.48e-05	0.582
Benign vs Malignant (Correct Non-Focus)				
Fractal Dimension	0.031	-0.304	0.352	-0.00434
Lacunarity	0.0347	0.583	0.00173	0.638
Lbp	0.0197	-0.439	0.0478	-0.293
Glcm_contrast	0.00154	-0.432	0.0181	-0.374
Glcm_homogeneity	0.00258	0.338	0.0555	0.0904

Table 8: Full statistical test results for dependent (top 4 rows) and independent tests (bottom 4 rows) for every considered measure on both validation and test set for the 40x BreakHIS dataset.

Feature	P-value (test-set)	Cohen's D (test-set)	P-value (val-set)	Cohen's D (val-set)
Focus vs Non-focus (Malignant)				
Fractal Dimension	5.69e-08	0.421	2.63e-08	0.466
Lacunarity	1.04e-19	-1.03	2.19e-16	-0.855
Lbp	0.125	-0.0986	0.317	0.0152
Glcm_contrast	0.00622	0.184	0.367	0.115
Glcm_homogeneity	0.588	0.0506	0.459	-0.121
Focus vs Non-focus (Benign)				
Fractal Dimension	2.46e-05	0.614	0.000788	0.507
Lacunarity	4.01e-12	-2.38	6.36e-11	-2.47
Lbp	0.335	0.0814	0.176	0.283
Glcm_contrast	0.0261	0.297	0.103	0.214
Glcm_homogeneity	0.000955	-0.451	2.78e-05	-0.569
Focus vs Non-focus (Correct Malignant)				
Fractal Dimension	1.13e-05	0.388	8.44e-06	0.426
Lacunarity	6.05e-16	-0.979	5.41e-10	-0.695
Lbp	0.113	-0.0812	0.00514	-0.224
Glcm_contrast	0.0296	0.168	0.361	0.11
Glcm_homogeneity	0.17	0.118	0.694	0.0383
Focus vs Non-focus (Correct Benign)				
Fractal Dimension	2.45e-05	0.472	7.85e-05	0.478
Lacunarity	6.91e-14	-1.5	7.49e-12	-1.14
Lbp	0.422	0.00319	0.612	0.177
Glcm_contrast	0.194	0.106	0.379	0.0843
Glcm_homogeneity	0.0108	-0.176	0.0136	-0.297
Benign vs Malignant (Focus)				
Fractal Dimension	0.823	-0.0544	0.435	0.0959
Lacunarity	0.0264	-0.31	0.321	-0.0599
Lbp	0.743	0.121	0.214	0.233
Glcm_contrast	0.74	-0.102	0.0531	0.274
Glcm_homogeneity	0.442	0.0617	0.00446	-0.423
Benign vs Malignant (Non-Focus)				
Fractal Dimension	0.227	-0.179	0.0803	0.155
Lacunarity	0.00264	0.448	0.0687	0.313
Lbp	0.0728	-0.0117	0.699	-0.0189
Glcm_contrast	0.442	-0.0759	0.0771	0.286
Glcm_homogeneity	0.00541	0.355	0.0608	-0.119
Benign vs Malignant (Correct Focus)				
Fractal Dimension	0.833	0.0217	0.483	0.114
Lacunarity	0.0417	-0.344	0.259	-0.0661
Lbp	0.673	0.201	0.0326	0.358
Glcm_contrast	0.31	0.0652	0.0245	0.348
Glcm_homogeneity	0.352	-0.18	0.000126	-0.655
Benign vs Malignant (Correct Non-Focus)				
Fractal Dimension	0.245	-0.237	0.0836	0.108
Lacunarity	3.04e-06	0.883	1.08e-09	1.28
Lbp	0.168	0.0238	0.0728	-0.315
Glcm_contrast	0.258	-0.154	0.368	0.222
Glcm_homogeneity	0.0128	0.423	0.575	0.135

Table 9: Full statistical test results for dependent (top 4 rows) and independent tests (bottom 4 rows) for every considered measure on both validation and test set for the 400x BreakHIS dataset.