



UPPSALA  
UNIVERSITET

**Is eXplainable AI suitable as a hypotheses generating tool for  
medical research? Comparing basic pathology annotation  
with heat maps to find out**

Albert Adlersson

*Bachelor's thesis in Statistics*

*Advisor*

Patrik Andersson

# Abstract

Hypothesis testing has long been a formal and standardized process. Hypothesis generation, on the other hand, remains largely informal. This thesis assess whether eXplainable AI (XAI) can aid in the standardization of hypothesis generation through its utilization as a hypothesis generating tool for medical research. We produce XAI heat maps for a Convolutional Neural Network (CNN) trained to classify Microsatellite Instability (MSI) in colon and gastric cancer with four different XAI methods: *Guided Backpropagation*, *VarGrad*, *Grad-CAM* and *Sobol Attribution*. We then compare these heat maps with pathology annotations in order to look for differences to turn into new hypotheses. Our CNN successfully generates non-random XAI heat maps whilst achieving a validation accuracy of 85% and a validation AUC of 93% – as compared to others who achieve a AUC of 87%. Our results conclude that Guided Backpropagation and VarGrad are better at explaining high-level image features whereas Grad-CAM and Sobol Attribution are better at explaining low-level ones. This makes the two groups of XAI methods good complements to each other. Images of Microsatellite Instability (MSI) with high differentiation are more difficult to analyse regardless of which XAI is used, probably due to exhibiting less regularity. Regardless of this drawback, our assessment is that XAI can be used as a useful hypotheses generating tool for research in medicine. Our results indicate that our CNN utilizes the same features as our basic pathology annotations when classifying MSI – with some additional features of basic pathology missing – features which we successfully are able to generate new hypotheses with.

**Keywords:** black box, eXplainable AI (XAI), Convolutional Neural Network (CNN), Microsatellite Instability (MSI), colon cancer, gastric cancer, hypotheses generating, hypotheses generating tool, medical research

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Data</b>	<b>3</b>
<b>3</b>	<b>Theory</b>	<b>4</b>
3.1	eXplainable AI . . . . .	4
3.2	eXplainable AI instrument selection . . . . .	7
3.2.1	Guided Backpropagation . . . . .	8
3.2.2	VarGrad . . . . .	9
3.2.3	Grad-CAM . . . . .	10
3.2.4	Sobol Attribution . . . . .	11
3.3	Basic Pathology . . . . .	12
3.4	Basic Pathology Annotation selection . . . . .	13
3.4.1	Tumor Infiltrating Lymphocytes . . . . .	15
3.4.2	Chron's like reaction . . . . .	16
3.4.3	Poor differentiation . . . . .	16
3.4.4	Mucinous or signet ring cell morphology . . . . .	17
<b>4</b>	<b>Method</b>	<b>19</b>
4.1	Creating the Convolutional Neural Network . . . . .	19
4.2	Training the Convolutional Neural Network . . . . .	20
4.3	Evaluating the Convolutional Neural Network . . . . .	20
<b>5</b>	<b>Results</b>	<b>22</b>

5.1	Result of model parameter randomization test . . . . .	22
5.1.1	Test-results for Guided Backpropagation . . . . .	23
5.1.2	Test-results for VarGrad . . . . .	23
5.1.3	Test-results for Grad-CAM . . . . .	24
5.1.4	Test-results for Sobol Attribution . . . . .	24
5.2	Result of comparison between Basic Pathology Annotation (BPA) and eX- plainable AI . . . . .	25
5.2.1	Comparison between BPA and Guided Backpropagation . . . . .	26
5.2.2	Comparison between BPA and VarGrad . . . . .	27
5.2.3	Comparison between BPA and Grad-CAM . . . . .	28
5.2.4	Comparison between BPA and Sobol Attribution . . . . .	29
<b>6</b>	<b>Discussion</b>	<b>30</b>
<b>7</b>	<b>Conclusion</b>	<b>31</b>
<b>8</b>	<b>Acknowledgements</b>	
<b>9</b>	<b>Remarks</b>	
<b>10</b>	<b>Appendix</b>	
10.1	Images selected for analysis in Section 5 . . . . .	
10.2	Architecture of Convolutional Neural Network . . . . .	
10.3	Results of Basic Pathology Annotation . . . . .	
10.4	Detailed results of comparison between Basic Pathology Annotation and eX- plainable AI . . . . .	

# 1 Introduction

Along with the rise of information technology has come a sophistication in medical data formats – such as the 1985 development of the Digital Imaging and Communications in Medicine (DICOM) standard of communication, DICOM (2023). Since then, a mounting supply of data has been made available for use with Artificial Intelligence (AI): The Cancer Imaging Archive (TCIA) and The Cancer Genome Atlas (TCGA) as two prominent examples, Koh et al. (2022); Dlamini et al. (2020). A use which has taken its expression in the automation of the diagnosis of diseases in many medical fields, dermatology and mammography just to name two, Esteva et al. (2017). Together with the increasing demand of AI comes an mounting appeal for more transparent AI models – in contrast to the "black-box" systems that take inputs to generate outputs, with the in-between veiled by a black box. Users press AI designers for more than answers through predictions and now require inference as well – the difference between prediction and inference, akin to the difference between 'a weather forecast' and 'an explanation for cloud formation and how variations in air pressure affect precipitation'. The drawback of AI as compared to a physician is that it lacks the ability to provide explanations for its reasoning whereas a physician can refer to medical theory or job experience. In order to provide users with inference, one must learn to open the AI's black box. Opening an AI's black box is possible, but requires certain tools.

The foremost toolbox for explaining AI is called eXplainable AI (XAI) and contains many methods, all specialising in different angles of explanation. The current demand for XAI mainly comes from healthcare, law and science. Healthcare often utilize AI-generated decision trees to better convey and motivate why a patient is recommended a specific treatment. Law requires similar explainability tools to ensure that algorithm generated credit scores do not break the law by declining loans to borrowers on basis of sex or ethnicity, Hickling et al. (2023). Lastly, the area which could stand to benefit the most from an increase in explainability – is science. XAI has the potential to aid scientists in better understanding their scientific models. Something which could aid them in improving them further, furthering our understanding of our world, Lindskog and Ljung (1994); Arain et al. (2012); Ma et al. (2023); Greydanus et al. (2019); He and Yang (2020); Forssell and Lindskog (1997).

This thesis will consider XAI's application in medical science. Hypothesis testing has long been a formal and standardized process. Hypothesis generation, on the other hand, remains

largely informal. The purpose of this thesis is to assess whether XAI can formalise and standardize this process in order to make medical research more efficient and accessible for digital pathologists and statisticians alike. In order to further exemplify what we mean with "hypothesis generating tool", we bring up an example from Ludwig and Mullainathan (2023). In their working paper, titled *Machine Learning as a Tool for Hypothesis Generation*, they illustrate a procedure where they utilize a hypothesis generation tool for generating hypothesis for what could be the reason for a judge's decisions about who to jail. Interestingly they are able to generate hypothesis that the defendant's face matters surprisingly much for the judge's decision – but not due to demographics nor existing psychology research, hence the difference necessary for generating new hypothesis.

Hickling et al. (2023) find it difficult to draw conclusions about which XAI methods are best for medical applications – as they only came across one such application. Their study recommends further research into the medical field by applying some of the XAI methods described in their paper. This thesis will follow their recommendation by utilizing XAI to explore an AI's focus areas in images of Microsatellite Instability (MSI). Identifying MSI in cancer patients is important since it determines their response towards immunotherapy. We will compare the XAI heat maps to our own annotations for visual features important in basic pathology related to MSI. Finally, we compare how the two identification models - the XAI and basic pathology - differ in their predictions and discuss hypothetical reasons as to why. The purpose of this thesis is to try to generate hypotheses from from these unexplained differences for future medical research to explore. This thesis attempts to answer the following question: **"Is eXplainable AI suitable as a hypotheses generating tool for medical research?"**

The thesis is organised as follows: In Section 2 we go over details for our data set, alongside our motivations for selecting it. In Section 3 we go over how the thesis applies eXplainable AI and basic pathology in a way necessary for us to investigate the thesis question. In Section 4 we create, train and evaluating our Convolutional Neural Network (CNN) for classifying Microsatellite Instability. In Section 5, we present the result of a model parameter randomization test as well as the result of our Basic Pathology Annotation (BPA) and eXplainable AI comparison – divided into one subsection for each of our four selected XAI instruments. In Section 6 we discuss the result in Section 5. Lastly, in Section 7, we conclude the findings of our thesis – alongside recommendations for future studies.

## 2 Data

This thesis utilizes a data set downloaded from Kaggle (2019). The data set is based on a data set from *Zenodo* authored by Kather (2019). It contains 192 312 unique image patches derived from histological images of colorectal and gastric cancer patients. Schirris et al. (2022) state that the Kather data set has been collected from 360 patients (The Cancer Genome Atlas - Colorectal Carcinoma). These image patches have then been divided into the following two classes: 'MSS' (Microsatellite Stable or "healthy" image patches) and 'MSIMUT' (Microsatellite Instable or highly Mutated, "sick" image patches). Microsatellite Instability is a harmful condition which can occur in a number of places within the body. Due to our data being of colon and gastric cancer only, we will limit our thesis to the study of MSI in these places. Out of the 192 312 unique images available in the Kather data set, we utilize 150 078.

The focus of this thesis is on the analysis of medical image data derived from Formalin-Fixed Paraffin-Embedded (FFPE) slides. FFPE slides are the standard for diagnostic medicine and generated by fixing a specimen in formaldehyde and then inserting it into a paraffin wax block for cutting – giving it a well preserved appearance suitable for computational analysis. Most of The Cancer Genome Atlas images are of frozen specimens and thus not suitable for computational analysis. Flash frozen samples frequently damages the tissue. Since we utilize images derived from FFPE slides, we should not experience this problem. The FFPE slides which lie as the basis of our image patches have the following preprocessing applied to them: automatic detection of tumor, resizing to 224 px x 224 px at a resolution of 0.5  $\mu\text{m}/\text{px}$ , color normalization with the Macenko method – Macenko et al. (2009) – and assignment of patients to either 'MSS' or 'MSIMUT'. Matek et al. (2021) utilizes eXplainable AI in their paper exploring differentiation of bone marrow cell morphologies. They present XAI heat maps of a 224 px x 224 px resolution. Judging by these heat maps fuzzy, low defined appearance, we would caution against basic pathology interpretations of images below this resolution.

It is important that the digital pathology images are of high enough resolution in order to make the resulting eXplainable AI heat maps interpretable. If the image patches sampled from the slides are too narrow in scope and size, the AI will be left with an area too minor for developing an understanding of where in the colon or gastrointestinal tract the tissue was sampled from. If the image patches sampled from the slides are too broad in scope and

size, the AI will be left with relatively fewer image patches to train on – thus increasing the risk that the AI fails to reach an accuracy that is satisfactory. In a way, the patch size is a balancing act between under- and over-fitting our AI model, a problem much akin to the variance-bias trade-off. Ideally, we would want somewhat higher of a resolution than the 224 px x 224 px resolution available to us via our selected data set.

Schirris et al. (2022) use the same dataset as us. Although they also write about MSI prediction for colorectal and gastric cancer, their paper focuses on feature extraction and modelling tumor heterogeneity. They do not focus on eXplainable AI as a hypotheses generating tool for medical research, but instead write about a Deep learning-based weak label learning method for analyzing Whole Slide Images (WSIs). They also utilize WSIs whereas we utilize WSIs divided into patches.

### 3 Theory

This section is divided into four sections. In Section 3.1, we go over how eXplainable AI can aid medical research. In Section 3.2, we continue with our motivation for selecting our four XAI instruments, along with how each of these function. In Section 3.3, we briefly explain Microsatellite Instability. Lastly, in Section 3.4, we motivate our selection of visual features for our basic pathology annotation.

#### 3.1 eXplainable AI

As was explained in the introduction of this thesis, users now require inference from their AI and not only predictions. The underlying factors motivating this demand differs depending on the end user. In the context of a patient, inference could be in the shape of an explanation motivating their recommended treatment – making them feel safer due to the increase in trust recommendations, with the arguments attached, have on a patient. In the context of a medical researcher – as is the scenario in this thesis – inference could take the shape of an explanation of which features an AI deem most important whilst arriving at its conclusion. The difference between an AI prediction and inference can be exemplified by the difference between receiving a prediction that a patient is infected with a particular virus, and receiving an explanation of how this virus causes harm to the patient – information important for the development of a vaccine.

Although the prospect of increased AI inference seem positive, it is not without its potential drawbacks. There is currently a debate over whether an AI models degree of explainability compromises the models precision and effectiveness. One side of the debate emphasise the value gained from simplifying models in order to make them easier to interpret – an argument in line with the scientific principle of parsimony. The other side of the debate instead emphasise the value gained from added abstraction – stemming from preserving the AI models complexity. Their side argues that it is impossible to keep an optimized accuracy and precision whilst also keeping the model understandable, Ghassemi et al. (2021). Regardless of which side one supports, the liveliness of the debate indicates the continued importance of inference and explainability in the context of AI models.

Adebayo et al. (2018) suggest utilizing a test in order to test whether our XAI is succeeding in visualizing the inner workings of the model or the data generating process – thus providing us with more information of which side of the debate our particular AI model leans toward. One of these tests is the *model parameter randomization test*. The model parameter randomization test compares the eXplainable AI output from the AI model before it is trained with the output after it is trained. This in order to investigate whether the XAI method is sensitive to the properties of the AI model or not. If the XAI method is sensitive to the learned parameters, we should expect the XAI output from the untrained and trained AI to differ substantially. If the XAI outputs are very similar, we probably have an indication that the XAI method does not capture the AI models underlying procedure. The result of this test is presented in Section 5 of this thesis. In the case of our thesis, it is performed by a test comparing the XAI output of the complete 25 epoch trained AI model with the XAI output of a 1 epoch trained AI model.

Savage provide one example of a eXplainable AI medical research application, Nature (2022). In the early days of the COVID-19 pandemic, when radiographs of COVID-19 infected people were scarce, the scarcity of radiographs led scientists to complement their data sets with radiographs of healthy people from the US National Institutes of Health (NIH). These radiographs contained systematic differences – unrelated to the disease – from the COVID-19 radiographs. Radiographs usually label a person’s right side with the letter ‘R’ in the top corner of each X-ray. With most of the images of healthy people stemming from a single source, some of the AI systems based their diagnoses on the style and placement of the letter ‘R’, rather than on the outward state of the lungs – see Figure 1.

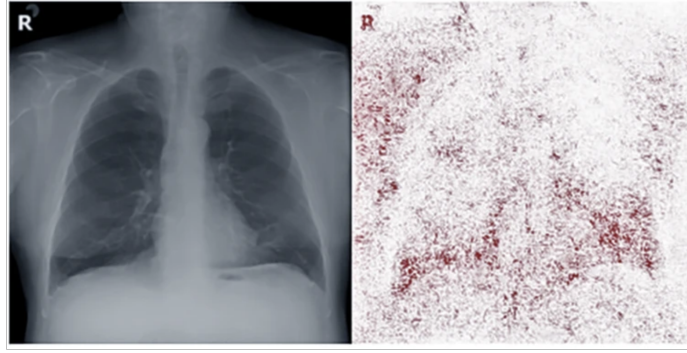


Figure 1: The part of the X-ray that AI were using to diagnose COVID-19 (red pixels) included the letters that marked patients’ right sides, Bustos et al. (2020).

What made XAI useful was its ability to visualise this flaw in the AI’s classification model so that scientists could calibrate their AI to better capture the data it was really meant to model. This example proves how insight into AI’s inner workings can be valuable in a medical research context, since it helps scientists locate flaws in their models – with the added bonus that it could visualise patterns scientists did not previously know about, Nature (2022). It is the discovery of unknown patterns in medical image data that this thesis attempts to generate, so as to create new hypotheses for medical research.

In this thesis we will train a Convolutional Neural Network on a data set of histological images for Microsatellite Instability identification in gastrointestinal cancer. A Convolutional Neural Network (CNN) is a type of Artificial Intelligence designed to analyse images. It is composed of several stacks of layers – beginning with an input layer and ending in an output layer – which form a mathematical construct capable of identifying information from images it has been trained to recognize. It trains on labeled data – through supervised learning – and fine tunes its weights every time it gets the label wrong, until it arrives at a high enough prediction accuracy for it to end its training. A CNN identifies the hidden label of an image by extracting features from the image – features the CNN extracts via its utilization of convolutional, pooling and fully connected layers – which it then uses to make its prediction.

In clinical practice, not every patient is tested for MSI. Good diagnosis require additional genetic or immunohistochemical tests. This additional process for diagnosis takes valuable time from patients, potentially post-poning their treatment. Great sensitivity to visual image identification of MSI therefore has the potential to save valuable time and money. We therefore believe visual diagnosis of MSI to be a valuable medical research area to use as our testing ground for the eXplainable AI hypotheses generating method assessed by this thesis.

### 3.2 eXplainable AI instrument selection

The eXplainable AI toolbox contains many instruments. It is useful to know which instruments best fit which kind of situation. Vilone and Longo (2021) divide XAI into five general categories based on the XAI's explanation format output. The five categories are: *numeric*, *rule based*, *textual*, *visual* and *mixed* explanation. They further state four factors to consider whilst selecting a suitable XAI instrument for one's application. The four factors are: *methods for explainability*, *field of application*, *types of users*, and lastly *purposes of explanation*. We will analyze how these four factors relate to the purposes of our AI and assess which of the five categories of XAI we should utilize.

First, since we are using image data, it makes sense to utilize an image based instrument. This in order to keep the interpretation of the output as intuitive as possible. Second, our field of application is a rather narrow field of academia, so universal interpretability will not be as important. Third, our intended users are digital pathologists and medical researchers, so we want a method that provides inference into our AI's inner workings. Instilling trust in our users is thus secondary to obtaining good inference. Fourth, since pathologists and medical researchers might lack knowledge about AI, it is important that we utilize an XAI method that is as intuitive as possible – thus allowing for our medical specialist to apply their knowledge without obstruction. Fifth, the purpose of the explanation is knowledge discovery via the generation of new hypotheses for medical research.

It is important to consider the pros and cons of each of the five categories of XAI methods before deciding on one to utilize. Numeric XAI output is flexible but not as intuitive, rule based schematic output is structured but difficult to scale, text based output is intuitive but long and drawn out, image based output is informative but dependent on outside information like legends and captions to be correctly interpreted, lastly mixed output is a compromise well suited for groups with varying end users – yet complex and confusing if sizeable enough. Keeping the pros and cons of each category in mind, we decide upon selecting the image category for our XAI. Partly due to Savage's previous example, Nature (2022).

Proceeding with an image-based XAI, one attempts to suggest diseases which are possible to identify through images – in contrast to diseases only identifiable through blood samples or immunohistochemical tests. When considering different diseases to identify, one might reason that some diseases are better captured by images than others. A sick body part with

a large three-dimensional volume, like a pair of lungs or a brain, might be less consistent with a two-dimensional image than say a birthmark – which is relatively flat. That said, we do find papers with XAI implementations for the three dimensional organs of the body. For example, for brain related diseases such as Alzheimer, Parkinson’s and Schizophrenia, Bloch and Friedrich (2022). Despite this, we believe it best to stick to our analysis of relatively ”two dimensional” biopsy slide image data.

Our thesis utilizes the *Xplique* library for applying XAI, Xplique (2023a). The code for implementing the XAI is based on example code from the same library, Xplique (2023b). The different XAI methods we utilize are all image based, with heat maps as their output format. Heat maps visualise the most important areas for image identification in an input image – with important areas painted in ”warmer” colours (reds) and unimportant areas painted in ”colder” colours (blues). The library contains the following feature attribution methods for generating heat maps: *Saliency*, *Gradient Input*, *Guided Backprop*, *Integrated Gradients*, *Smooth Grad*, *Square Grad*, *VarGrad*, *Grad-CAM*, *Occlusion*, *Rise* and *Sobol Attribution*. Due to time limitations, not making the thesis too lengthy, and due to the fact that many of the different methods produce very similar results – we decided upon selecting only four of these methods for analysis in our thesis. We tried all of these methods on a random sample from our data set to see how each respective method’s heat map differed. The four methods that differed most, whilst also remaining interpretable, and which we ended up in our thesis, are: *Guided Backpropagation*, *VarGrad*, *Grad-CAM* and the *Sobol Attribution method*. These XAI methods are presented and explained in Section 3.2.1, Section 3.2.2, Section 3.2.3 and Section 3.2.4 respectively. What makes these methods differ from each other is the way in which they arrive at this heat map.

### 3.2.1 Guided Backpropagation

Guided backpropagation generates its XAI heat map by combining a *backpropagation function* with a *backward ’deconvnet’ function*, Springenberg et al. (2015). *Backpropagation* is an algorithm for updating weights used to calibrate a Neural Network’s decision making. The function applies an input vector to the network and then propagates forward from the input layer to the output layer. An error value is calculated by taking the desired output minus the actual output for each of the network output neurons. The error value is propagated backward as a function of the contribution of the error when accounting for the network weights. This

organizes the network such that the hidden layer recognizes features in the input space. The output layer is then able to use the hidden layer features to arrive at a solution, IBM (2017). The *backward 'deconvnet' function* refers to the backward 'deconvolutional neural network' function. It uses deconvolutional layers to upsample the image so that it can generate feature visualizations. The function applies deconvolutional layers in reverse, during the backpropagation phase of the neural network. It visualises concepts learned in the high-level layers of the CNN by using a high-level feature map and inverting the CNN's data flow – going from neuron activation's in the given layer down to an image. Typically, a single neuron is left non-zero in the high level feature map. Then, the resulting reconstructed image shows the part of the input image that is most strongly activating this neuron, Springenberg et al. (2015).

*Guided backpropagation* combines the *backpropagation* and *backward 'deconvnet'* function to visualize the activation of high layer neurons. Given an input image, it performs the forward pass to the layer we are interested in. It then sets to zero all activation's except for one, and propagates back to the image to obtain a reconstruction. The formula for guided backpropagation is,

$$(f_i^l > 0) \times (R_i^{l+1} > 0) \times R_i^{l+1} = R_i^l$$

where  $(f_i^l > 0)$  is the backpropagation part with the network activation's and  $(R_i^{l+1} > 0)$  is the backward 'deconvnet' part with the network gradients. The gradients quantify how much a change in each input dimension change the predictions around the input.  $f_i^0$  is the input image before the forward pass,  $f_i^L$  is the input image after the forward pass,  $R_i^0$  is the reconstructed image after the backward pass and  $R_i^L$  is the reconstructed image before the backward pass. Guided backpropagation aims to zero out negative gradients during computation of 'intermediate representations' obtained during the backward pass. It does this by only keeping the positive activation's and gradients. It is important to note that the 'deconvnet' approach and guided backpropagation do not compute a true gradient but an imputed one, Springenberg et al. (2015).

### 3.2.2 VarGrad

VarGrad generates its XAI heat map through a variance analog of SmoothGrad. SmoothGrad averages over explanations of noisy copies of an input by drawing noise vectors  $g_i \sim N(0, \sigma^2)$  i.i.d. from a normal distribution, Adebayo et al. (2018). VarGrad is independent of the

gradient and is able to capture higher order partial derivatives. The formula for VarGrad is,

$$E_{vg}(x) = V(E(x + g_i)),$$

where  $V$  corresponds to the variance. VarGrad is an estimator of the gradient of the Kullback-Leibler divergence. The Kullback-Leibler divergence is a statistical distance and a measure of how one probability distribution is different from a reference probability distribution. It is interpreted as the average difference of the number of bits required for encoding samples of one probability distribution utilizing a code optimized for a reference probability distribution, Saltelli (1951). VarGrad is an unbiased estimator of the gradient of this divergence and based on Reinforce with leave-one-out control variables. It utilizes a score function method as an estimator for Variational Inference. The goal of Variational Inference is to approximate the posterior distribution of a model. Variational Inference accomplishes this by utilizing a parameterised family of distributions, finding the parameters by minimising the Kullback-Leibler divergence. This estimator is then connected to the log-variance loss which is defined as the variance of the log ratio – which has the property of reproducing the gradients of the Kullback-Leibler divergence under certain conditions.

Since the Kullback-Leibler divergence is intractable, Variational Inference can cast the inference problem as an optimisation problem – a problem which can be solved with stochastic optimisation tools. In particular, Variational Inference forms a Monte Carlo estimator of the gradient of the Evidence Lower Bound (ELBO). VarGrad can be implemented via an algorithm structured in the following way. First, we sample from the approximate posterior. Second, we detach the samples from the computational graph. Third, we get an estimate of the negative ELBO and the log-variance loss. Finally, we differentiate through the loss with respect to our variational parameter of choice, Richter et al. (2020).

### 3.2.3 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) generates its XAI heat map by utilizing the gradient information in the last convolutional layers of the CNN. It does this in order to compromise between high-level semantics and detailed spatial information. Grad-CAM assigns importance values to each neuron for a particular decision of interest in order to explain activation’s in any selected layer of a neural network. In order to obtain the class-discriminative localization map for any class, it first computes the gradient score for our selected class with respect to feature map activation’s of a convolutional layer. These

gradients are then global-average-pooled to obtain the neuron weights ( $\alpha_k^c$ ), Selvaraju et al. (2019). The neuron weights ( $\alpha_k^c$ ) are defined as,

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \times \frac{\partial y^c}{\partial A_{ij}^k}.$$

where ( $\alpha_k^c$ ) represents a *partial linearization* of the deep network, downstream from  $A^k$  which are the feature map activations of a convolutional layer, and captures the ‘importance’ of feature map  $k$  for target class  $c$ . We then perform a weighted combination of forward activation maps, and follow it by a *ReLU* activation function to calculate,

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k).$$

The result is a coarse heat map of the same size as the convolutional feature map. The reason why we apply a ReLU to the linear combination of maps is because we are interested in the features that have a *positive* influence on the class of interest. Negative pixels likely belong to other categories than those we are interested in, Selvaraju et al. (2019).

### 3.2.4 Sobol Attribution

Sobol Attribution generates its XAI heat map by taking the variance between the input and output of a CNN to produce prediction scores for the selected layers. The individual pixel importance scores used to paint the heat map are calculated by sampling masks from a Quasi-Monte Carlo sequence. After these masks have been sampled, they are applied to the input image through a perturbation function – a function which modifies input in order to observe the change in output. This forms stochastic random perturbed inputs that then are forwarded into the CNN we want to obtain prediction scores for, FEL et al. (2021). Lastly, the prediction scores are calculated by adapting Sobol-based sensitivity analysis (also referred to as Variance-based sensitivity analysis).

Sobol-based sensitivity analysis is a form of *global* sensitivity analysis – sensitivity analysis being the study of how the uncertainty in the output of a model can be attributed to different sources of uncertainty in the input. Global sensitivity analysis is simply a category of sensitivity analysis which measures sensitivity across the whole input space (*i.e.* global). The analysis decomposes the variance of the output of the model into fractions which can be attributed to inputs. These percentages are directly interpreted as measures of sensitivity. Variance-based measures of sensitivity are useful because they can deal with nonlinear

responses and measure the effect of interactions in non-additive systems, Saltelli and Annoni (2010).

### 3.3 Basic Pathology

Pathology is defined as the study of the cause and effect of disease or injury. It is a complex field of medicine which consider visual features of specimen in order to identify disease. To narrow the scope of this thesis, we will only make use of basic visual features when attempting to identify Microsatellite Instability. For this thesis, we will limit our scope to the annotation of MSI occurring within gastrointestinal cancer – that is colorectal and gastric cancer. Gastrointestinal cancer tumors commonly express MSI. That being said, MSI is prevalent in many different cancers. Knowledge of the prevalence of MSI is important since it can affect treatment options and prognosis of patients.

Microsatellite Instability is a genetic disease which occurs when an individual accumulates mutations in DNA regions known as microsatellites – microsatellites being regions of repeated DNA that show instability. MSI mainly occurs due to one of two reasons. It either occurs due to Replication Errors (RER) or it occurs due to chromosomal instability (CIN). RER occurs due to faults with the mismatch repair (MMR) system, whereas CIN occurs due to a broader instability involving larger segments of DNA, as well as chromosomes. MMR is a system for recognizing and repairing incorrect incorporation of bases, and can arise during DNA replication and recombination. The most important genes for a functioning mismatch repair system are, in order of importance: MSH2, MLH1, MSH6 and PMS2, YouTube (2020).

Due to the genetic nature of MSI, pathology currently relies on genetic or immunohistochemical tests to identify those affected. One need to stain the specimen with pigments to determine whether the four genes (MSH2, MLH1, MSH6 and PMS2) in the tumors are functioning. That is, if the tumor cells turn brown (positive) or blue (negative) – with brown (positive) being an indication of active MSI – when stained, YouTube (2020). If one could discover new visual patterns for classifying MSI, a lot of time and effort spent on genetic and immunohistochemical tests could be spared. In this thesis, we hope to generate new hypotheses related to new visual features for MSI.

### 3.4 Basic Pathology Annotation selection

Currently known visual features indicative of MSI include, but are not limited to: *Tumor Infiltrating Lymphocytes* (TILs), *Chron's like reaction*, *poor differentiation* and lastly *mucinous or signet ring cell morphology*. The reason why we have decided on considering these particular features is that these are the subset of the set of features visual during microscopy which also do not require any additional information on where in the body the specimen has been sampled from – information which we can not easily implement to our CNN, YouTube (2020). It are these basic visual features which we will consider and annotate for the images we have selected to analyse in Section 5. Section 3.4.1, Section 3.4.2, Section 3.4.3 and Section 3.4.4 each describe one of our four selected visual features and their respective appearance in microscopy imaging. *Note*, that these images are high resolution examples to illustrate the pathological features we are considering, and thus not images from our data set – which has a much lower 224 px x 224 px image resolution. Before we describe these four visual features in more detail, we provide an example of our annotation process. We do this by utilizing an image sampled from our data set. Figure 2 presents an example of our annotation process for Image 4 from Section 5.

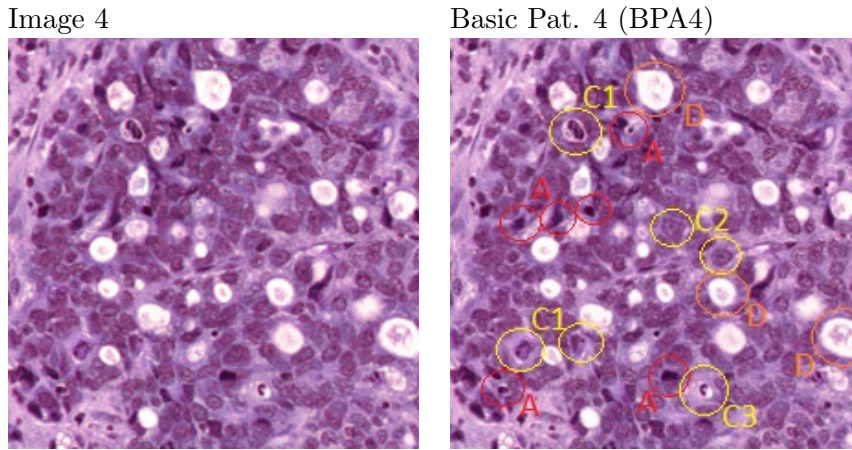


Figure 2: Annotation example for Image 4. 224 px x 224 px resolution.

In Figure 2, we first see a raw image as it is stored in our data set (left image). To annotate the image (Image 4), we copy it and begin to circle features from our list of visual signs of MSI. We do not circle every example of each feature – one example per feature present is enough. The visual basic pathology visual features are annotated as: *Tumor Infiltrating*

*Lymphocytes* (TILs) [Annotated with a red 'A'], *Chron's like reaction* [Annotated with a blue 'B'], *poor differentiation* [Annotated with a yellow 'C'] and lastly *mucinous or signet ring cell morphology* [Annotated with an orange 'D']. We provide an excerpt from the Appendix with the Basic Pathology Annotation (BPA) for Image 4 to further exemplify the results of (BPA4) – that is the second image in Figure 2 (right image).

## 1. Image 4 – Basic Pathology Annotation 4 (BPA4):

### (a) *Tumor Infiltrating Lymphocytes*

*Image 4* seem to contain a few more immune cells compared to what we saw in the MSS-class images (*Image 1-3*). These are annotated in red with an 'A' next to them. The relatively high quantity of TILs could be indicative of the image belonging to the MSIMUT-class.

### (b) *Chron's like reaction*

### (c) *Poor differentiation*

Does all of the cell look like surrounding cells? [No, the cells and glands appear different in shape and colour.] The height of the cell: cylinder, cuboids, slice etc? [No, the cells and glands appear to differ in height.] Nucleus shape and colour? [The resolution of the image is low, but from what is discernible there seems to be different sized nuclei, indicative of nuclear atypia or pleomorphism. These are annotated in yellow with a 'C1' next to them.] Chromatin packing or density? [The resolution of the image is low, but from what is discernible there seems to be no signs of abnormal chromatin packing.] Nucleus position in cell and in relation to other cells? [Different cells have different nucleus position, so what is normal for one type of cell can be abnormal for another. This makes our image difficult to interpret. The resolution of the image is low, but from what is discernible the visible nuclei seems to have somewhat differently positioned nuclei. Compare nuclei annotated yellow and with a 'C1' next to them with those annotated with a 'C2' next to them.] Quantity of mitosis? [The resolution of the image is low, only two possible mitosis is discernible. See top yellow 'C1' ring annotation and bottom yellow 'C3' annotation.] All things considered, *Image 4* appears to have a rather high differentiation, thus indicating that it belongs to the 'MSIMUT' class.

### (d) *Mucinous or signet ring cell morphology*

Mucinous or signet ring cell morphology is only visible in very aggressive types and stages of cancer. *Image 4* seem to possibly exhibit some signs of mucinous or signet ring cell morphology, although it is somewhat difficult to tell due to the low resolution and image patch scope – regardless of this, the possible mucinous or signet ring cell morphology is indicative of the image belonging to the 'MSIMUT' class. See orange annotated with 'D' next to them for examples.

In order to not bias our basic pathology annotation, we annotate the six randomly selected images presented in Section 5 before we look at the XAI heat maps – this in order to insure that our CNN's areas of importance do not influence our basic pathology annotation. The file names of the six images alongside the un-annotated images are presented in Table 3 in Section 10.1 of the Appendix. Due to space limitations, we will only include the Basic Pathology Annotations (BPA) – for the six images analyzes in Section 5 – in Section 10.3 of the Appendix. We now continue to describe the four visual features indicative of MSI which we have selected to annotate for our six images in Section 5. We do so in Section 3.4.1, Section 3.4.2, Section 3.4.3 and Section 3.4.4.

### 3.4.1 Tumor Infiltrating Lymfocytes

Observations of *Tumor Infiltrating Lymfocytes* (TILs) is our first visual indication of MSI. TILs are suggestive of Microsatellite Instability and may be seen in Lynch syndrome. The most important genes for a functioning mismatch repair system are, in order of importance: MSH2, MLH1, MSH6 and PMS2, YouTube (2020). These genes are called the 'Lynch syndrome genes' after the syndrome that appears in an individual when the genes are not working properly, CDC (2018).

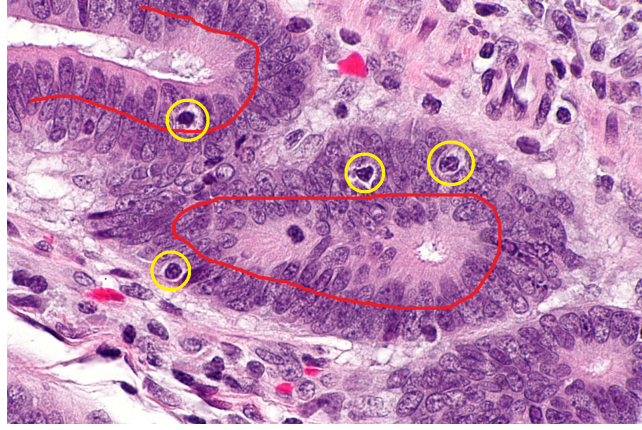


Figure 3: Micrograph with Hematoxylin and Eosin stain (H&E stain) showing tumor-infiltrating lymphocytes in colorectal carcinoma, Wikipedia (2013).

Figure 3 presents an image of TILs in colorectal carcinoma. The tumor glands in Figure 3 are annotated with red. The nuclei that exhibit a retraction artefact are T-cells and annotated with yellow. These T-cells are called Tumor Infiltrating Lymphocytes since they lie in-between the tumor glands and thus infiltrate them. When we believe that we have found TILs in our image, we annotate these by encircling them in a red ellipsoid, placing the letter 'A' adjacent to our observation.

### 3.4.2 Chron's like reaction

Observations of *Chron's like reaction* is our second visual indication of MSI. Chron's like reaction is difficult to evaluate since the images in our data set only provides us with a small window of the colon or gastric tissue. We do not have information on whether the images are of colon or gastric samples so we can not determine whether our images exhibit Chron's like reaction or not. Because of these limitations, we will not be able to use Chron's like reaction as a visual criteria for MSI. If we would have been able to include Chron's like reaction as a visual indication of MSI, it would have been annotated by encircling observations in a blue ellipsoid, placing the letter 'B' adjacent to our observation.

### 3.4.3 Poor differentiation

Observations of *poor differentiation* is our third visual indication of MSI. Differentiation describes the processes by which immature cells become mature cells and how similar the tumor tissue looks like the normal tissue it originated from. Well-differentiated cancer cells

look more like normal cells and tend to grow and spread more slowly than poorly differentiated cancer cells. Differentiation is used in tumor grading systems and are different for each cancer NIH (2023).

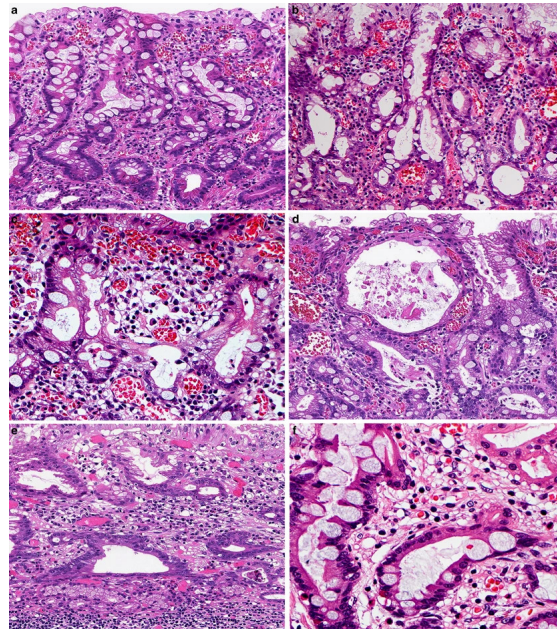


Figure 4: Visualisation example of poorly differentiated tissues. (a) Representative examples of tortuous glands, (b) a branching gland, (c) anastomosing glands, (d) a distended gland, (e) spiky glands, (f) and glandular outgrowth in very well-differentiated adenocarcinoma of intestinal type, Ushiku et al. (2013).

Figure 4 demonstrates poorly differentiated cancerous gastric tissues in various locations of the intestine. When assessing the level of differentiation of our image, we will also look for features with answers the following questions: *Does the cell look like surrounding cells?*, *What is the height of the cell, is it cylindrical or cuboid or slice like?*, *What is the nucleus shape and colour?*, *Is there any signs of chromatin packing?*, *What is the nucleus position in the cells in relation to other cells?* and lastly *What is the quantity of mitosis?*

When we believe that we have found indications of poor differentiation in our image, we annotate these by encircling them in a yellow ellipsoid, placing the letter 'C' adjacent to our observation.

#### 3.4.4 Mucinous or signet ring cell morphology

Observations of *Mucinous or signet ring cell morphology* is our fourth and last visual indication of MSI. Signet ring cell carcinoma (SRCC) and mucinous adenocarcinoma (MCC) are

histologic subtypes of colon adenocarcinoma. Signet ring cell cancers are most commonly seen in the stomach (95%) and occasionally found in colon, rectum, ovary, peritoneum and gallbladder. It is characterized by specific morphologic appearance of abundant intracytoplasmic mucin pushing nucleus to the periphery giving it a signet ring cell appearance, Thota et al. (2013).

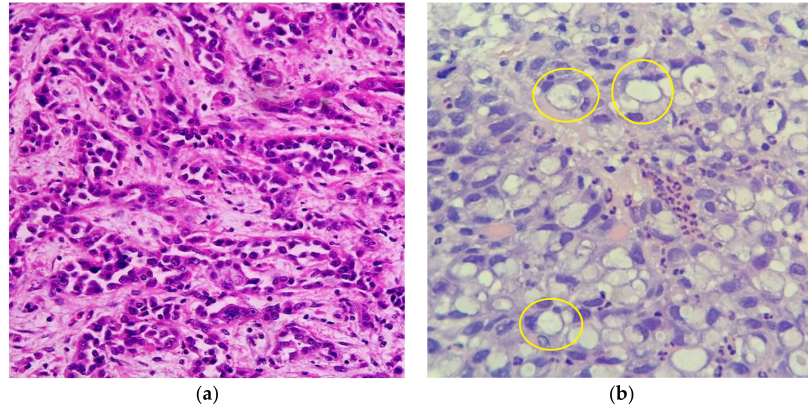


Figure 5: (a) poorly differentiated conventional gastric adenocarcinoma, Commons (2005). (b) gastric signet adenocarcinoma, Commons (2017).

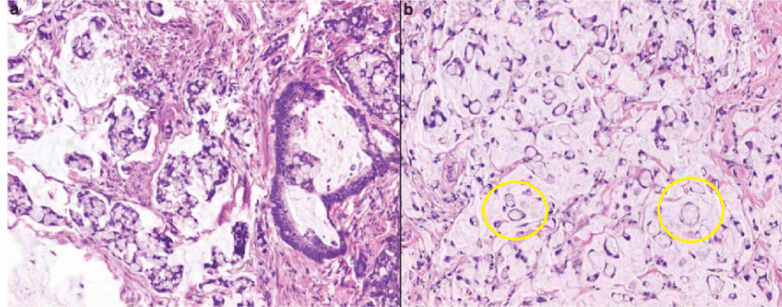


Figure 6: (a) High-grade mucinous adenocarcinoma with signet ring cells composed of mixed mucinous (right half of image) and signet ring cell components. (b) High-grade mucinous adenocarcinoma with signet ring cells composed predominantly of neoplastic signet ring cells associated with copious extracellular mucin, Davison et al. (2014).

Figure 5 and Figure 6 both demonstrate *mucinous or signet ring cell morphology*. The areas annotated in yellow are the signet ring cells. When we believe that we have found indications of mucinous or signet ring cell morphology in our image, we annotate these by encircling them in a orange ellipsoid, placing the letter 'D' adjescent to our obsevation.

## 4 Method

This section is divided into three parts. In Section 4.1, we create our CNN. In Section 4.2, we train our CNN. Lastly, in Section 4.3, we evaluate our CNN.

### 4.1 Creating the Convolutional Neural Network

We utilize Python 3.9 for programming. Python is the main programming language for machine learning with *Keras* – an Application Programming Interface (API) for Tensorflow which makes programming AI more user-friendly. The CNN is based on code by Keras (2022) and built as an un-optimized version of the Xception network architecture. Xception is a deep Convolutional Neural Network architecture that involves Depthwise Separable Convolutions, introduced by Francois Chollet – the creator of Keras, IQ (2023). A simplified summary of the architecture of the CNN is presented in Table 1 below.

Layer type	Output Shape	Param #
InputLayer	None, 224, 224, 3	0
Rescaling	None, 224, 224, 3	0
Conv2D	None, 112, 112, 12, 8	3584
BatchNormalization	None, 112, 112, 12, 8	512
Activation	None, 112, 112, 12, 8	0
Activation + SeparableConv2D + Batch-Normalization + Activation + Separable-Conv2D + BatchNormalization + Max-Pooling2D + Conv2D + Add	Input shape: None, 112, 112, 12, 8 Output shape: None, 56, 56, 256	Input params: 0 Output params: 0
Activation + SeparableConv2D + Batch-Normalization + Activation + Separable-Conv2D + BatchNormalization + Max-Pooling2D + Conv2D + Add	Input shape: None, 56, 56, 256 Output shape: None, 28, 28, 512	Input params: 0 Output params: 0
Activation + SeparableConv2D + Batch-Normalization + Activation + Separable-Conv2D + BatchNormalization + Max-Pooling2D + Conv2D + Add	Input shape: None, 28, 28, 512 Output shape: None, 14, 14, 728	Input params: 0 Output params: 0
SeparableConv2D	None, 14, 14, 1024	753048
BatchNormalization	None, 14, 14, 1024	4096
Activation	None, 14, 14, 1024	0
GlobalAveragePooling2D	None, 1024	0
Dropout	None, 1024	0
Dropout	None, 1024	0
Dense	None, 1	1025

Table 1: Summary of the Convolutional Neural Network architecture.

The CNN architecture is composed of a total of 2 731 065 parameters, where 2 722 777 are trainable and 8 288 non-trainable. Each row in Table 1 represents one layer type, or one group of layer types. The layers are sorted in their "chronological" order, with the first layer at row one, and the last layer at the last row as a layer of type "Dense". The "Output Shape" column holds information on the shape of the output of the layer, with the first

layer "InputLayer" having an input 224 px high and 224 px wide, with a three channel RGB colour depth. The last column "Param #" counts the number of parameters for each layer, or group of layers. For a more in depth view of the CNN architecture, see Section 10.2 in the Appendix.

## 4.2 Training the Convolutional Neural Network

The CNN is trained for 25 epochs, with batches of 16 images each. We use a subset of 150 078 images for utilization by the CNN. This data is then divided into training and validation-data, with a 0.2 validation split for each of the two partitions. Out of 150 078 images belonging to two classes, o model utilizes 120 063 images for training and 30 015 images for validation. To increase the out-of-sample accuracy of the CNN we incorporate two regularization methods – methods used to simplify and balance the complexity of a model. First, we incorporate data augmentation by randomly flipping our images horizontally and vertically, as well as rotating them. Then, we also incorporate several dropout layers.

## 4.3 Evaluating the Convolutional Neural Network

The metrics for the CNN, after having been trained for 25 epochs, are presented in Table 2. We chose to end the training process after 25 epochs in order to decrease the risk of over-fitting the CNN model. We also ended training at 25 epochs in order to save time, this since 25 epochs took 16 hours. The training loss and validation loss are presented in Figure 7. The loss of the CNN is a summation of the errors made from each sample in training or validation sets. The goal of the training process is to minimize the loss. Lower loss generally means a better performing model. We arrived at a validation loss of 0.4012 for our CNN – see Table 2. The accuracy and validation accuracy are also presented in Figure 7. The accuracy of the CNN is also a metric of model performance. It is the the count of predictions where the predicted value is equal to the true value. We arrived at a validation accuracy of 0.8544 for our CNN - see Table 2.

<b>Train loss</b> 0.1730	<b>Train accuracy</b> 0.9289	<b>Binary train accuracy</b> 0.9289	<b>Train AUC</b> 0.9816
<b>Validation loss</b> 0.4012	<b>Validation accuracy</b> 0.8544	<b>Binary validation accuracy</b> 0.8544	<b>Validation AUC</b> 0.9369

Table 2: CNN model evaluation metrics.

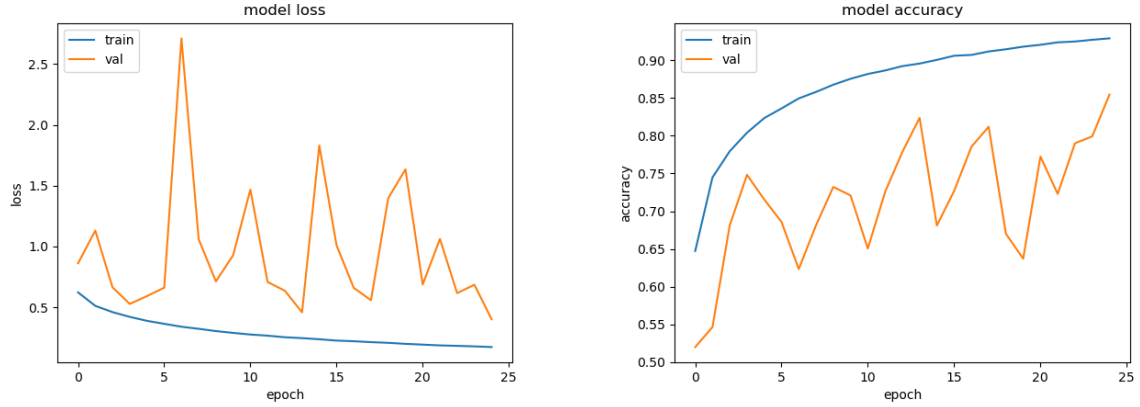


Figure 7: Metrics for each epoch – (Model loss to the left, Model accuracy to the right).

The high validation accuracy of ca 85% has been achieved in previous studies – Schirris et al. (2022) achieve an AUC of 87% whereas we achieve an AUC of 93% – but remains surprisingly good considering that pathology currently relies on a set of additional tests to identify MSI patients. To really know whether a patient has MSI, one need to stain the specimen with pigments revealing whether the four genes MSH2, MLH1, MSH6 and PMS2 are working or not. Our data set is not stained for these genes and is thus at a disadvantage in comparison to genetic and immunohistological tests, Kather (2019). Despite these obstacles, we succeed in achieving a high validation accuracy of 85% – something which is rather surprising.

In a medical context, it is often of high importance to understand the difference between sensitivity and specificity. This due to the high stakes of medical decisions since they can mean the difference between unnecessary prolonged suffering and potential death or relief from ones disease. Sensitivity refers to a tests capability to identify a positive observation as positive whereas specificity refers to a tests capability to identify a negative observation as negative. The Area Under the Curve (AUC) is a metric that measures the CNN models ability to correctly identify positive observations as positive, and negative observations as negative. The theoretically highest possible AUC score is 1, and means that all observations where correctly identified. Since we achieve a validation AUC of 0.9369 – which is very close to 1 – this is a very good result.

## 5 Results

This section is divided into two parts. In Section 5.1, we present the results of our model parameter randomization test – this in order to see if our XAI really functions the way we wish. In Section 5.2, we present the result of the comparison between Basic Pathology Annotation (BPA) and eXplainable AI.

### 5.1 Result of model parameter randomization test

This section is divided into four sections – Section 5.1.1, Section 5.1.2, Section 5.1.3 and Section 5.1.4 – with each section being associated with one of our four selected XAI instruments. The top row of each figure, Figure 8-11, represents Image 1-3 – that is the 'MSS'-true classes (as in "healthy"). The bottom row of each figure, Figure 8-11, represents Image 4-6 – that is the 'MSIMUT'-true classes (as in "sick"). All images present in Figure 8-11 are of XAI heat map output for our CNN, fixed for training in only one epoch – as opposed to 25 epochs, which is the setting for our fully trained CNN (the results of which are presented in Section 5.2).

Judging by the result of the model parameter randomization test – as indicated by Figure 8, Figure 9, Figure 10 and Figure 11 – it seems like our four selected XAI methods do capture the AI's underlying identification process, Adebayo et al. (2018). This based on the observation that our XAI heat maps – as presented in Section 5.2 – exhibit much more structure relative to the heat maps presented in this section, Section 5.1. We note that 'Grad-CAM 2' in Figure 11, Section 5.1.3, failed to render.

### 5.1.1 Test-results for Guided Backpropagation

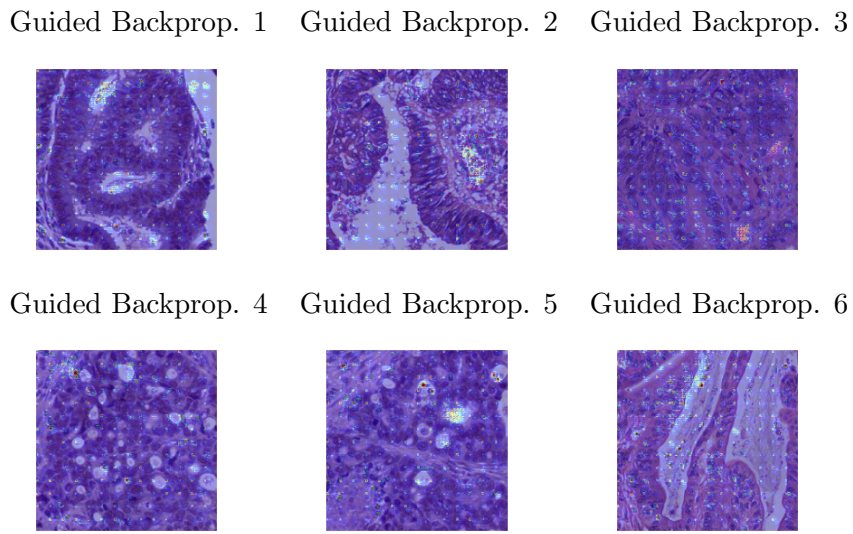


Figure 8: **Test results, Row 1-2: 'MSS' = TRUE ; Test results, Row 3-4: 'MSIMUT' = TRUE**

### 5.1.2 Test-results for VarGrad

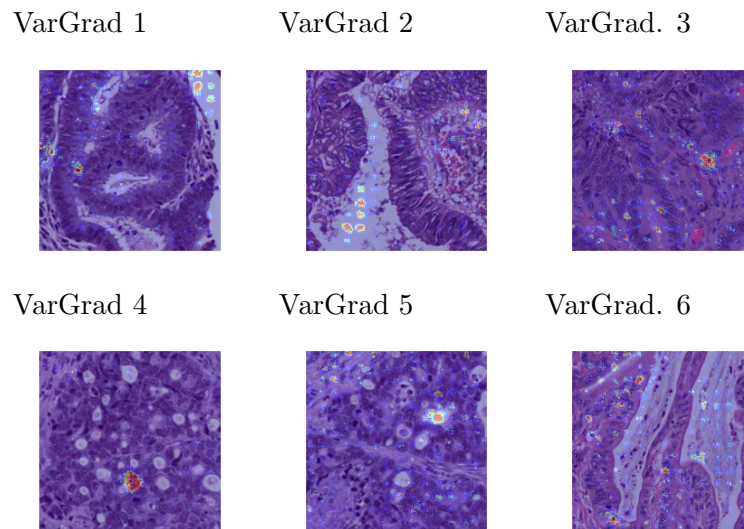


Figure 9: **Test results, Row 1-2: 'MSS' = TRUE ; Test results, Row 3-4: 'MSIMUT' = TRUE**

### 5.1.3 Test-results for Grad-CAM

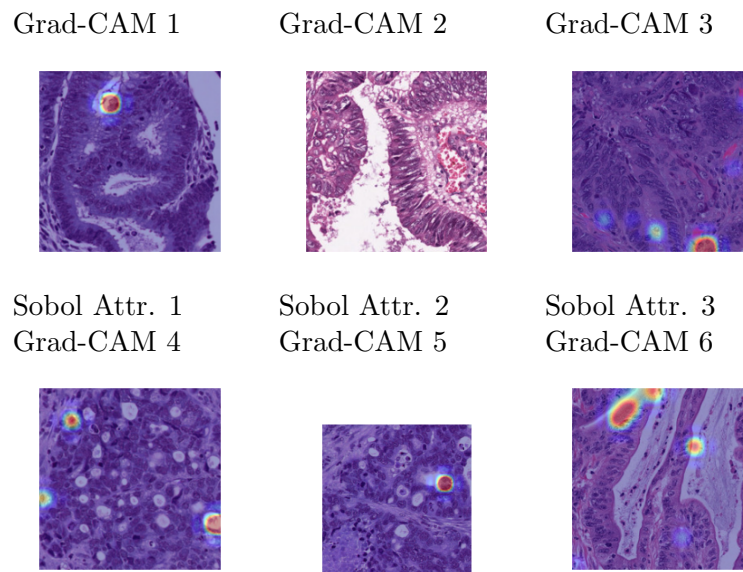


Figure 10: **Test results, Row 1-2: 'MSS' = TRUE ; Test results, Row 3-4: 'MSIMUT' = TRUE**

### 5.1.4 Test-results for Sobol Attribution

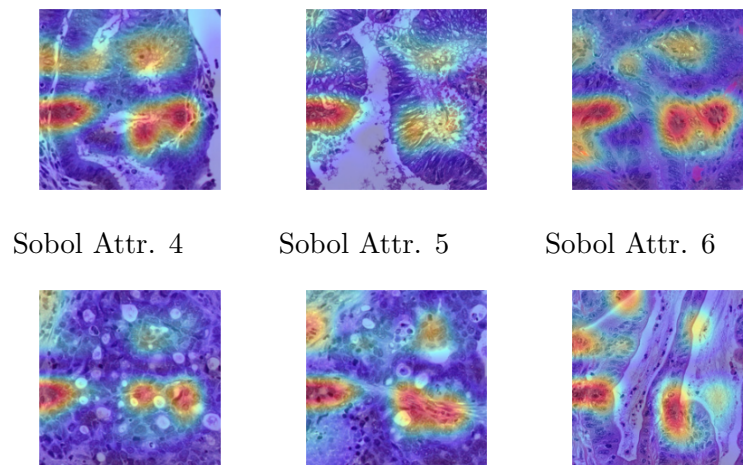


Figure 11: **Test results, Row 1-2: 'MSS' = TRUE ; Test results, Row 3-4: 'MSIMUT' = TRUE**

## 5.2 Result of comparison between Basic Pathology Annotation (BPA) and eXplainable AI

This section is divided into four sections – Section 5.2.1, Section 5.2.2, Section 5.2.3 and Section 5.2.4 – with each section being associated with one of our four selected XAI instruments. The top two rows of each figure, Figure 12-15, represents Image 1-3 – that is the 'MSS'-true classes (as in "healthy"). The bottom two rows of each figure, Figure 12-15, represents Image 4-6 – that is the 'MSIMUT'-true classes (as in "sick"). All images present in Figure 12-15 are of XAI heat map output for our CNN, fixed for our fully trained CNN – the one trained for 25 epochs – as opposed to the one epoch setting of our model parameter randomization test in Section 5.1. For detailed notes of Basic Pathology Annotation and XAI output comparison – on an image by image basis – see Section 10.4 in the Appendix.

### 5.2.1 Comparison between BPA and Guided Backpropagation

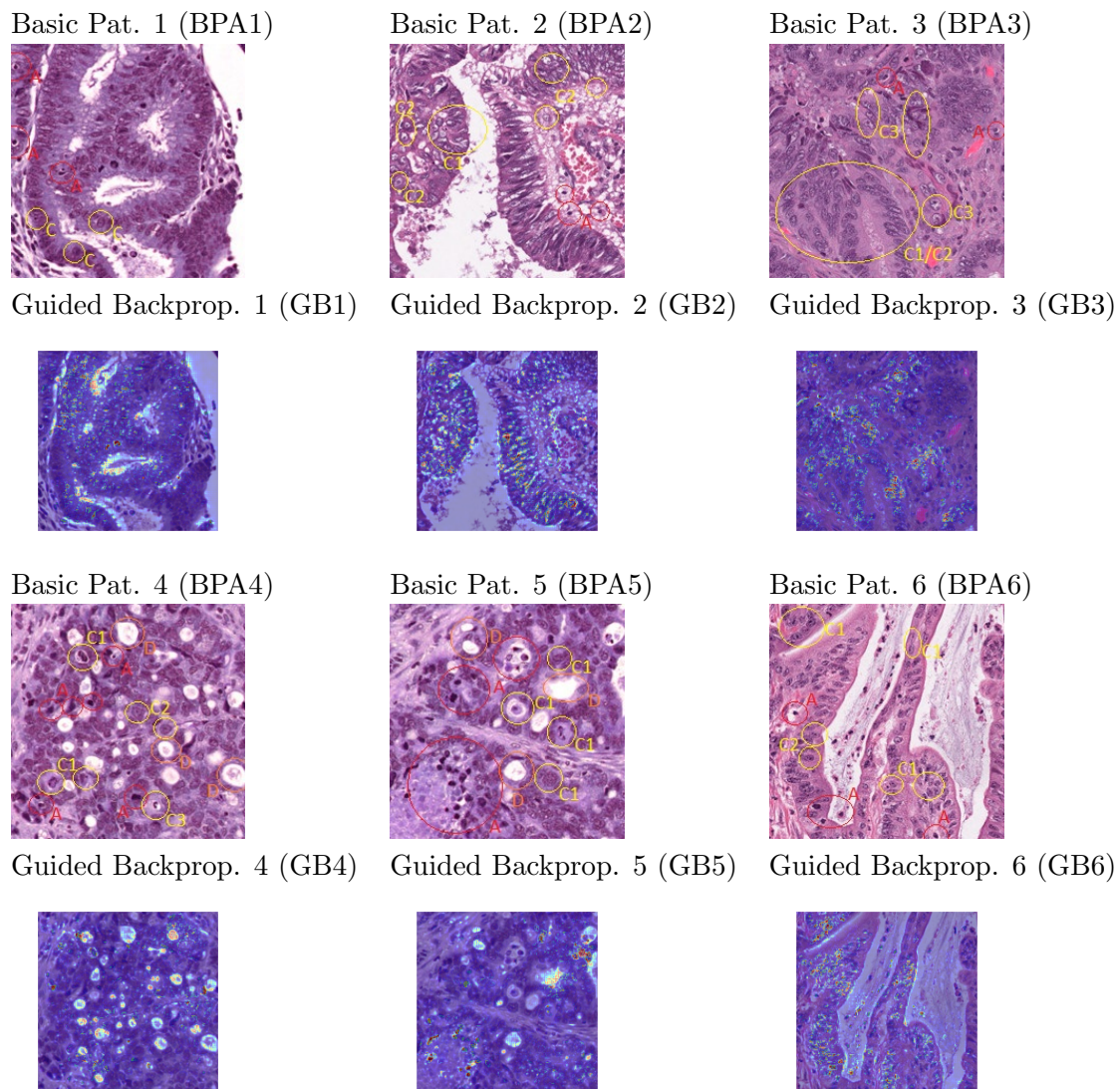


Figure 12: **Row 1-2:** 'MSS' = TRUE ; **Row 3-4:** 'MSIMUT' = TRUE

*Guided Backpropagation* (GB1-6) did not mark every Infiltrating Lymphocyte that the basic pathology annotation did – but marked other ILs instead. Guided Backpropagation was generally better at agreeing with basic pathology when taking 'MSIMUT'-class images as input, as compared to 'MSS'-class images. Guided Backpropagation marked most of the 'C's (signs of poor differentiation) when taking 'MSS'-class images as input. Guided Backpropagation searched for cell nuclei and signs of mitosis. It also searched for furrows, wrinkles and grooves. Guided Backpropagation seems to engage in differentiation feature detection.

### 5.2.2 Comparison between BPA and VarGrad

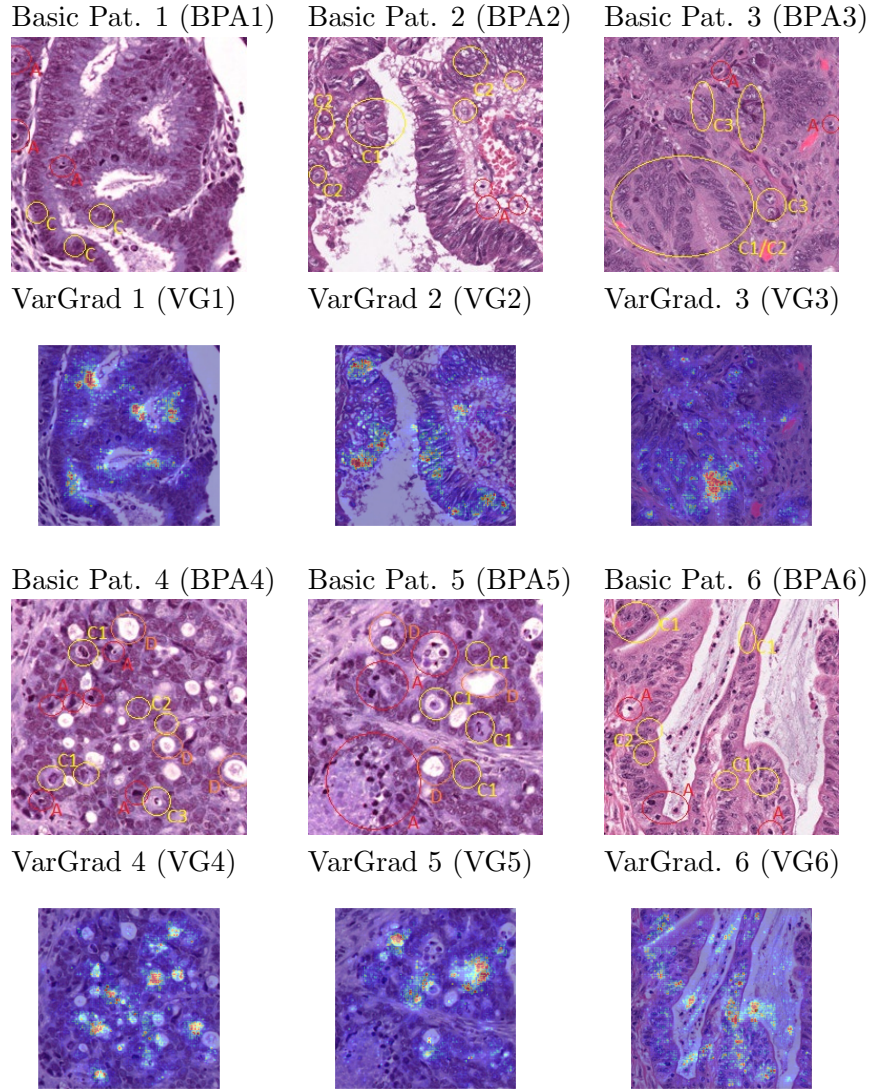


Figure 13: **Row 1-2:** 'MSS' = TRUE ; **Row 3-4:** 'MSIMUT' = TRUE

*VarGrad* (VG1-6) emphasized the white inside of the glandular architecture. *VarGrad* was somewhat less spread out in its focus areas in comparison to the other three XAI instruments. *VarGrad* mainly focused on the faded, hard to spot, blurred ring cells when considering images of 'MSIMUT'-class. *VarGrad* images were sometimes difficult to interpret. *VarGrad* images seem somewhat confused by large areas of white, perhaps wrongly mistaking white areas for 'MSIMUT' ring cells.

### 5.2.3 Comparison between BPA and Grad-CAM

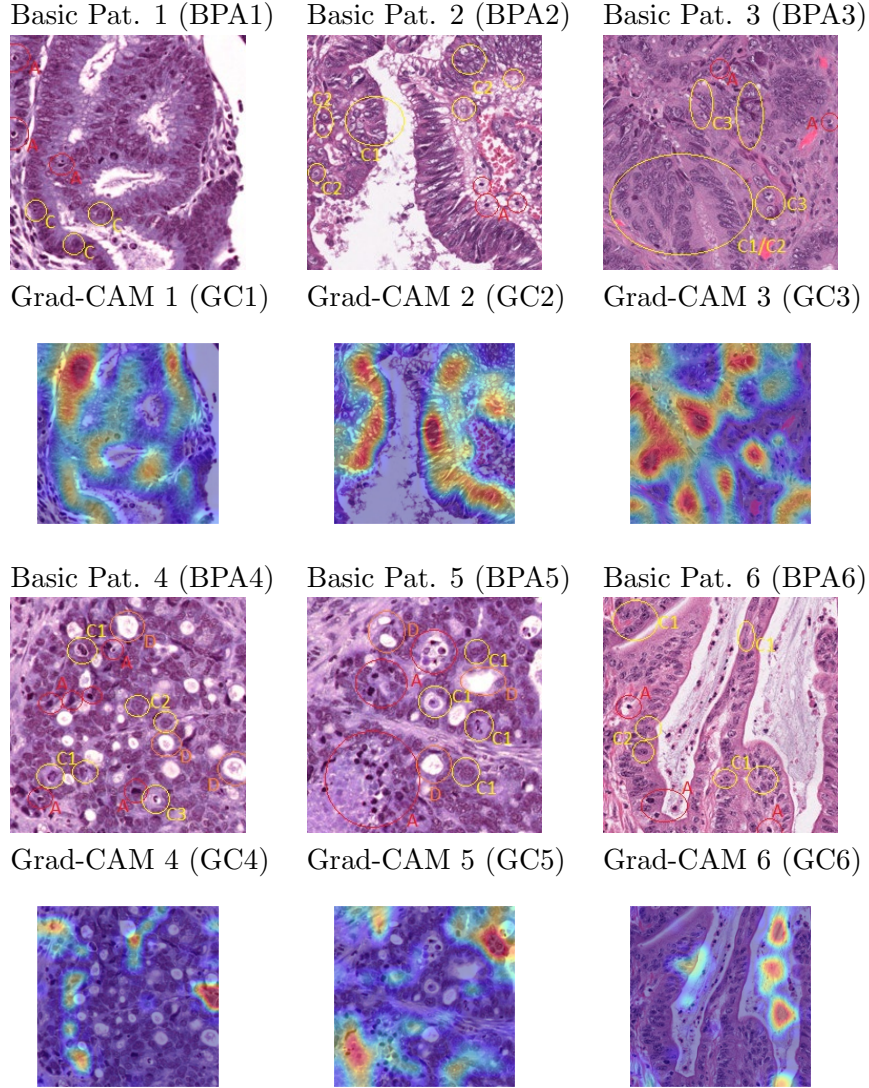


Figure 14: **Row 1-2:** 'MSS' = TRUE ; **Row 3-4:** 'MSIMUT' = TRUE

*Grad-CAM (GC1-6)* - emphasised the glands themselves, and their contribution to the overall architecture of the larger glandular tissue. It emphasised how they curved and formed. Grad-CAM seems to engage in differentiation feature detection, but struggles more in forming interpretable patterns of tissue architecture than Guided Backpropagation and VarGrad. Grad-CAM images are sometimes difficult to interpret – especially so when taking 'MSIMUT'-class images as input. Possibly because of the lessened level of regularity present in 'MSIMUT'-class images which often have poor differentiation. Grad-CAM images seem somewhat confused

by large areas of white, perhaps wrongly mistaking white areas for 'MSIMUT' ring cells. Grad-CAM seem to focus on "wrinkles" and "grooves" in the whitish tissue, perhaps to find regularity. Grad-CAM considers areas with LIs.

#### 5.2.4 Comparison between BPA and Sobol Attribution

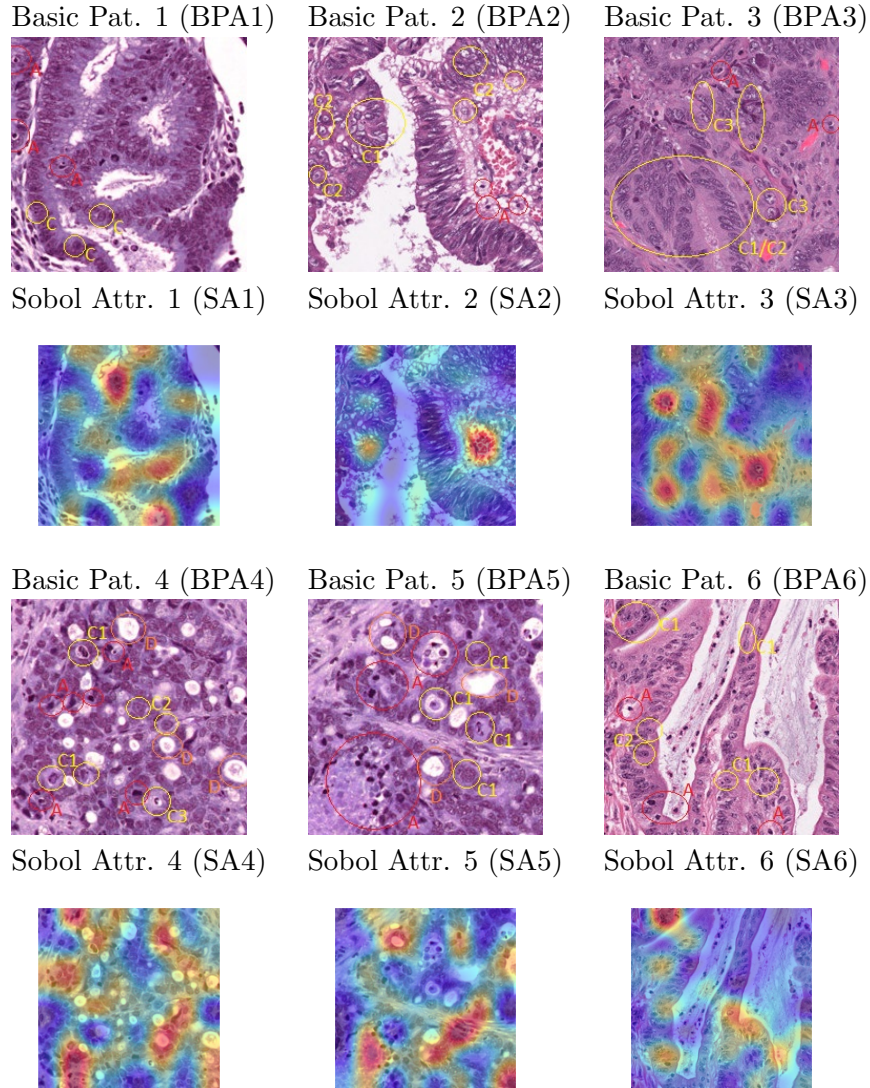


Figure 15: **Row 1-2:** 'MSS' = TRUE ; **Row 3-4:** 'MSIMUT' = TRUE

*Sobol Attribution (SA1-6)* - highlighted areas rather than details or features. The Sobol Attribution images were all difficult to interpret. Sobol Attribution seem to consider ILs and cell nuclei. Lastly, Sobol Attribution focused on erythrocytes (red blood cells) in SA2 – an area interestingly left out from BPA2.

## 6 Discussion

Through comparison between our Basic Pathology Annotations (BPA) and our XAI heat maps, we arrive at the following differences between the two:

Guided Backpropagation and VarGrad are generally better at explaining high-level features such as cell placement, components and structure of singular cells. Grad-CAM and Sobol Attribution are generally better at explaining low-level features such as tissue architecture and level of differentiation. We are unsure as to why these two groups of XAI instruments differ in their output, but we believe that it might have something to do with the instruments differing mathematical and technical construction. Regardless of why these XAI instruments differ, their differing abilities make them good complements.

Guided Backpropagation successfully considered Infiltrating Lymphocytes – but oftentimes other ILs than those initially annotated in our Basic Pathology Annotation. Guided Backpropagation also corresponded more with our Basic Pathology Annotation when the input image was of the 'MSIMUT'-class – as opposed to the 'MSS'-class. Guided Backpropagation marked more 'C's (signs of poor differentiation) when the input image was of the 'MSS'-class – as opposed to the 'MSIMUT'-class. Guided Backpropagation produced the most interpretable XAI heat maps. It is possible that Guided Backpropagation was less reliant on regularity in its image input – as opposed to the remaining three instruments – when attempting to visualize differentiation patterns in poorly differentiated tissue. We hypothesise that high-level features require higher resolution and image scope in order to be visualised correctly. Especially so for the 'MSIMUT'-class images, given their usually poor level of differentiation.

VarGrad and Grad-CAM both emphasized the whiter parts of the image. Both XAI instruments tended to focus more on faded, hard to spot ring cells – when the input image belonged to the 'MSIMUT'-class – than more well defined ring cells. Both instruments seemed potentially confused by images containing large amounts of white. Possibly due to wrongly mistaking these areas for 'MSIMUT' ring cells.

All four XAI methods seemed to consider ILs, cell nuclei and signs of mitosis. Guided Backpropagation and Grad-CAM also seemed to search for regularity through "furrows", "wrinkles" and "grooves" in addition to that. Lastly, Sobol Attribution is interesting since it

seems to focus on erythrocytes (red blood cells) in SA2 – an area interestingly left out from BPA2.

## 7 Conclusion

This thesis assess whether eXplainable AI can generate hypotheses for future medical research. It does so by creating AI generated heat maps to compare basic pathology annotation with. The purpose of this comparison is to look for differences between the AI’s identification for a disease and the basic pathology theory of identification for that same disease. Differences which we attempt to generate hypotheses from – hypotheses which hopefully will be interesting for future medical research.

This thesis produces XAI heat maps for a Convolutional Neural Network trained to classify Microsatellite Instability in colon and gastric cancer. It produces these utilizing four different XAI instruments: *Guided Backpropagation*, *VarGrad*, *Grad-CAM* and *Sobol Attribution*. Our CNN successfully passes the model parameter randomization test and successfully generates non-random XAI heat maps. It does so whilst achieving a validation accuracy of 85% and a validation AUC of 93% – as compared to Schirris et al. (2022) who achieve a AUC of 87%.

Out of the many observed differences between our basic pathology annotation and the XAI results, we are able to generate a multitude of hypotheses. For example: *Why did our CNN find blood vessels in SA2 important for classifying the tissue as 'MSS' (healthy)?*, *Why did the XAI seem to find blurred ring cells more indicative of 'MSIMUT' (sick) than less morphed ring cells?*, *Why did the XAI put greater emphasis on signs of poor differentiation than Tumor Infiltrating Lymphocytes?*. Since these hypotheses all seem worthy of investigation, we conclude that eXplainable AI can be successfully utilized as a hypotheses generating tool for generating hypotheses for medical research.

That said, there are some limitations to our study; mainly our lack of expert pathology annotations and low resolution image data set. Future studies should therefore use images of a higher resolution and scope in order to unlock the ability to analyse visual indications of MSI coupled to Chron’s like reaction, irregular tumor borders, infiltrative growth pattern and tumor necrosis – all features which our low-resolution image patches could not support. One suggestion for circumventing difficulties surrounding layman pathology assessment of low resolution medical image data is to utilize a numerical XAI method and numerical image

data which are not associated with the same problems of low resolution. We also recommend that future studies keep track of where in the body each image is sourced in order to further make use of additional visual indicators of MSI specific for these locations: gastric, colon, rectal and left or right side of the intestine.

It would be very interesting to do a follow up study in addition to this thesis where we utilize feature *visualisation* maps – as a complement to the feature *attribution* method results of this thesis – to develop a much deeper understanding for the basic pathology features most important for our Convolutional Neural Network’s classification of MSI. We recommend that a meta study be made of research which has utilized hypotheses generated from the study of XAI heat maps. It would be interesting to look into the results of those papers to further assess whether our method of generating hypotheses generally leads to significant results or not.

## 8 Acknowledgements

I want to thank my friend and classmate Alexander Koutakis for his support and advice and for reading my thesis and providing comments, my dear friend and corridor neighbour Alejandro Villaron who taught me some basics of coding, my corridor friend Marcus Liffler who let me borrow his book '*Basic Pathology, 10th edition – by Kumar, Abbas, Aster*' and taught me more about areas in medicine in need of time saving tools, my friend and corridor neighbour Hanna Rosendahl who taught me some basic pathology and let me borrow her book '*Junqueira’s Basic Histology - Text & Atlas - by Anthony L. Mescher*, Hampus Engström, Isak Åslund and Viktor Gånheim for reading my thesis and providing comments, and lastly my supervisor Patrik Andersson for providing advice and comments along the writing of the thesis. Big thanks to all who supported me during, and up until the completion, of this thesis.

## 9 Remarks

I am not a licensed pathologist or medical practitioner so take my annotations and application of medical theory with caution. Future studies would benefit from involving professional medical practitioners into the process of writing the thesis. First most in the annotation of

the images sampled for Section 5.

## References

- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B., 2018. Sanity checks for saliency maps, in: *Advances in Neural Information Processing Systems*, Curran Associates, Inc.
- Arain, M.A., Hultmann Ayala, H.V., Ansari, M.A., 2012. Nonlinear system identification using neural network, in: *Emerging Trends and Applications in Information Communication Technologies*, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 122–131.
- Bloch, L., Friedrich, C.M., 2022. Machine learning workflow to explain black-box models for early alzheimer’s disease classification evaluated for multiple datasets. *SN Computer Science* 3. doi:10.1007/s42979-022-01371-y.
- Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M., 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis* 66, 101797. doi:<https://doi.org/10.1016/j.media.2020.101797>.
- CDC, 2018. Microsatellite instability (msi) screening. URL: [https://www.cdc.gov/genomics/disease/colorectal\\_cancer/msi.htm](https://www.cdc.gov/genomics/disease/colorectal_cancer/msi.htm).
- Commons, W., 2005. File:adenocarcinoma low differentiated (stomach) he magn 400x.jpg. URL: [https://commons.wikimedia.org/wiki/File:Adenocarcinoma\\_low\\_differentiated\\_\(stomach\)\\_H%26E\\_magn\\_400x.jpg](https://commons.wikimedia.org/wiki/File:Adenocarcinoma_low_differentiated_(stomach)_H%26E_magn_400x.jpg).
- Commons, W., 2017. File:poorly cohesive gastric carcinoma (signet-ring cell type).jpg. URL: [https://commons.wikimedia.org/wiki/File:Poorly\\_cohesive\\_gastric\\_carcinoma\\_\(signet-ring\\_cell\\_type\).jpg](https://commons.wikimedia.org/wiki/File:Poorly_cohesive_gastric_carcinoma_(signet-ring_cell_type).jpg).
- Davison, J., Choudry, H., Pingpank, J., Ahrendt, S., Holtzman, M., Zureikat, A., Zeh, H., Ramalingam, L., Zhu, B., Nikiforova, M., Bartlett, D., Pai, R., 2014. Clinicopathologic and molecular analysis of disseminated appendiceal mucinous neoplasms: Identification of factors predicting survival and proposed criteria for a three-tiered assessment of tumor grade. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 27. doi:10.1038/modpathol.2014.37.

- DICOM, 2023. Dicom ps3.1 2023b - introduction and overview. URL: [https://dicom.nema.org/medical/dicom/current/output/chtml/part01/chapter\\_Foreword.html](https://dicom.nema.org/medical/dicom/current/output/chtml/part01/chapter_Foreword.html).
- Dlamini, Z., Francies, F.Z., Hull, R., Marima, R., 2020. Artificial intelligence (ai) and big data in cancer and precision oncology. *Computational and Structural Biotechnology Journal* 18, 2300–2311. doi:<https://doi.org/10.1016/j.csbj.2020.08.019>.
- Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542. doi:10.1038/nature21056.
- FEL, T., Cadene, R., Chalvidal, M., Cord, M., Vigouroux, D., Serre, T., 2021. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis, in: *Advances in Neural Information Processing Systems*, Curran Associates, Inc.. pp. 26005–26014.
- Forsell, U., Lindskog, P., 1997. Combining semi-physical and neural network modeling: An example of its usefulness. *IFAC Proceedings Volumes* 30, 767–770. doi:[https://doi.org/10.1016/S1474-6670\(17\)42938-7](https://doi.org/10.1016/S1474-6670(17)42938-7).
- Ghassemi, M., Oakden-Rayner, L., Beam, A.L., 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, e745–e750. doi:[https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9).
- Greydanus, S., Dzamba, M., Yosinski, J., 2019. Hamiltonian neural networks. [arXiv:1906.01563](https://arxiv.org/abs/1906.01563).
- He, F., Yang, Y., 2020. Nonlinear system identification of neural systems from neurophysiological signals. doi:10.1101/2020.08.09.243253.
- Hickling, T., Zenati, A., Aouf, N., Spencer, P., 2023. Explainability in deep reinforcement learning, a review into current methods and applications. [arXiv:2207.01911](https://arxiv.org/abs/2207.01911).
- IBM, 2017. A neural networks deep dive - an introduction to neural networks and their programming. URL: <https://developer.ibm.com/articles/cc-cognitive-neural-networks-deep-dive/>.
- IQ, O., 2023. Xception: Deep learning with depth-wise separable convolutions. URL: <https://iq.opengenus.org/xception-model>.

- Kaggle, 2019. Tcga coad msi vs mss prediction (jpg) - tiles of wsi of colorectal cancer (coad), ffpe samples. URL: [https://www.kaggle.com/datasets/joangibert/tcga\\_coad\\_msi\\_mss.jpg?resource=download](https://www.kaggle.com/datasets/joangibert/tcga_coad_msi_mss.jpg?resource=download).
- Kather, J.N., 2019. Zenodo – histological images for msi vs. mss classification in gastrointestinal cancer, ffpe samples [data set]., url = <https://doi.org/10.5281/zenodo.2530835>.
- Keras, 2022. Image classification from scratch. URL: [https://keras.io/examples/vision/image\\_classification\\_from\\_scratch/](https://keras.io/examples/vision/image_classification_from_scratch/).
- Koh, D.M., Papanikolaou, N., Bick, U., Illing, R., Kahn, C., Kalpathi-Cramer, J., Matos, C., Marti-Bonmati, L., Miles, A., Mun, S., Napel, S., Rockall, A., Sala, E., Strickland, N., Prior, F., 2022. Artificial intelligence and machine learning in cancer imaging. *Communications Medicine* 2, 133. doi:10.1038/s43856-022-00199-0.
- Lindskog, P., Ljung, L., 1994. Tools for semi-physical modeling. *IFAC Proceedings Volumes* 27, 1199–1204.
- Ludwig, J., Mullainathan, S., 2023. Machine learning as a tool for hypothesis generation. URL: <https://www.nber.org/papers/w31017>.
- Ma, D., Bortnik, J., Chu, X., Claudepierre, S.G., Ma, Q., Kellerman, A., 2023. Opening the black box of the radiation belt machine learning model. *Space Weather* 21. doi:10.1029/2022sw003339.
- Macenko, M., Niethammer, M., Marron, J., Borland, D., Woosley, J., Guan, X., Schmitt, C., Thomas, N., 2009. A method for normalizing histology slides for quantitative analysis., pp. 1107–1110. doi:10.1109/ISBI.2009.5193250.
- Matek, C., Krappe, S., Münzenmayer, C., Haferlach, T., Marr, C., 2021. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood* 138, 1917–1927. doi:<https://doi.org/10.1182/blood.2020010568>.
- Nature, 2022. Breaking into the black box of artificial intelligence. URL: <https://www-nature-com.ezproxy.its.uu.se/articles/d41586-022-00858-1>.
- NIH, 2023. differentiation. URL: <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/differentiation>.

- Richter, L., Boustati, A., Nüsken, N., Ruiz, F.J.R., Ömer Deniz Akyildiz, 2020. Vargrad: A low-variance gradient estimator for variational inference. [arXiv:2010.10436](https://arxiv.org/abs/2010.10436).
- Saltelli, A., Annoni, P., 2010. How to avoid a perfunctory sensitivity analysis. *Environmental Modelling Software* 25, 1508–1517. doi:<https://doi.org/10.1016/j.envsoft.2010.04.012>.
- Saltelli, A., A., 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 79 – 86. URL: <https://doi.org/10.1214/aoms/1177729694>, doi:10.1214/aoms/1177729694.
- Schirris, Y., Gavves, E., Nederlof, I., Horlings, H.M., Teuwen, J., 2022. Deepsmile: Contrastive self-supervised pre-training benefits msi and hrd classification directly from whole-slide images in colorectal and breast cancer. *Medical Image Analysis* 79, 102464. doi:<https://doi.org/10.1016/j.media.2022.102464>.
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2019. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* 128, 336–359. doi:10.1007/s11263-019-01228-7.
- Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2015. Striving for simplicity: The all convolutional net. [arXiv:1412.6806](https://arxiv.org/abs/1412.6806).
- Thota, R., Fang, X., Subbiah, S., 2013. Clinicopathological features and survival outcomes of primary signet ring cell and mucinous adenocarcinoma of colon: retrospective analysis of vaccr database. *Journal of Gastrointestinal Oncology* 5.
- Ushiku, T., Arnason, T., Ban, S., Hishima, T., Shimizu, M., Fukayama, M., Lauwers, G., 2013. Very well-differentiated gastric carcinoma of intestinal type: Analysis of diagnostic criteria. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 26. doi:10.1038/modpathol.2013.98.
- Vilone, G., Longo, L., 2021. Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction* 3, 615–661. doi:10.3390/make3030032.
- Wikipedia, 2013. File:tumour-infiltrating lymphocytes - 1 – very high mag.jpg. URL: [https://cs.m.wikipedia.org/wiki/Soubor:Tumour-infiltrating\\_lymphocytes\\_-\\_1\\_--\\_very\\_high\\_mag.jpg](https://cs.m.wikipedia.org/wiki/Soubor:Tumour-infiltrating_lymphocytes_-_1_--_very_high_mag.jpg).

Xplique, 2023a. Xplique – explainability toolbox for neural networks. URL: <https://deel-ai.github.io/xplique/>.

Xplique, G.C. 2023b. Welcome to the feature attribution tutorial. URL: <https://colab.research.google.com/drive/1XproaVxXjO9nrBSyyy7BuKJ1vy21iHs2>.

YouTube, 2020. Microsatellite instability and lynch syndrome. URL: <https://www.youtube.com/watch?v=5MWDGDvDDoo>.

## 10 Appendix

### 10.1 Images selected for analysis in Section 5

In order to not bias our basic pathology annotation, we annotate the six randomly selected images analyzed in the results of Section 5 before we look at the XAI heat maps – this in order to insure that our CNN’s areas of importance do not influence our basic pathology annotation. The file names of the six images alongside the un-annotated images are presented in Table 3. These same images are presented visually in Figure 16.

'MSS'=TRUE	'MSIMUT'=TRUE
<b>Jpg. 1 (No. 209):</b> blk-ACITVQNHAYKC-TCGA-AA-3818-01Z-00-DX1	<b>Jpg. 4 (No. 72 070):</b> blk-YFLKLYWGMPE-TCGA-AA-3811-01Z-00-DX1
<b>Jpg. 2 (No. 38 661):</b> blk-NAGCCGNAWYKC-TCGA-CA-5256-01Z-00-DX1	<b>Jpg. 5 (No. 7 765):</b> blk-DCCDNGAKKVSA-TCGA-AA-3811-01Z-00-DX1
<b>Jpg. 3 (No. 21 014):</b> blk-GYFFVKVWPHP-TCGA-CM-5864-01Z-00-DX1	<b>Jpg. 6 (No. 44 457):</b> blk-NVPATDVKDPTR-TCGA-CK-5913-01Z-00-DX1

Table 3: Randomly selected images to be analyzed in Section 5.

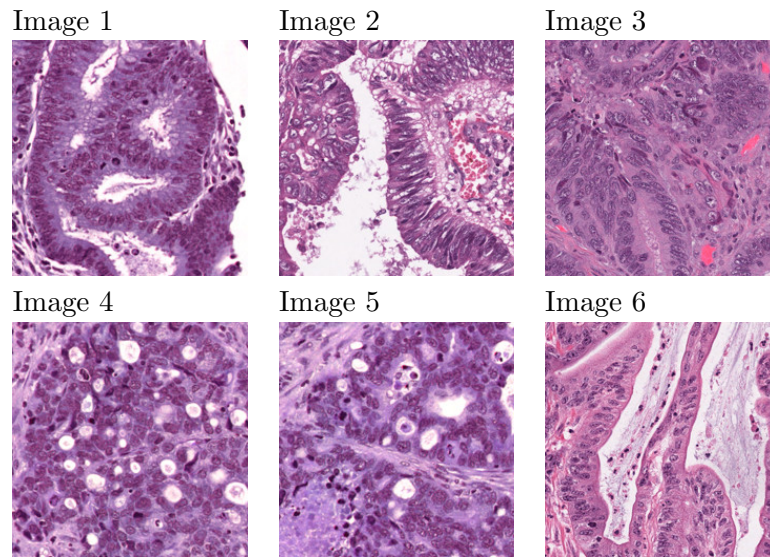


Figure 16: Non-annotated images (see Table 3) selected for analysis in Section 5.

## 10.2 Architecture of Convolutional Neural Network

A more detailed presentation of the architecture of our Convolutional Neural Network is presented in Table 4 and Figure 17. Table 4 and 17 hold the same information, the difference being that Figure 17 presents the information in Table 4 in a diagram format instead of in a table format.

Layer name	Layer type	Output Shape	Param #	Connected to
input_1	(InputLayer)	[(None, 224, 224, 3)]	0	[]
rescaling_1	(Rescaling)	(None, 224, 224, 3)	0	['input_1[0][0]']
conv2d	(Conv2D)	(None, 112, 112, 12, 8)	3584	['rescaling[0][0]']
batch_normalization	(BatchNormalization)	(None, 112, 112, 12, 8)	512	['conv2d[0][0]']
activation	(Activation)	(None, 112, 112, 12, 8)	0	['batch_normalization[0][0]']
activation_1	(Activation)	(None, 112, 112, 12, 8)	0	['activation[0][0]']
separable_conv2d	(SeparableConv2D)	(None, 112, 112, 25, 6)	34176	['activation_1[0][0]']
batch_normalization_1	(BatchNormalization)	(None, 112, 112, 25, 6)	1024	['separable_conv2d[0][0]']
activation_2	(Activation)	(None, 112, 112, 25, 6)	0	['batch_normalization_1[0][0]']
separable_conv2d_1	(SeparableConv2D)	(None, 112, 112, 25, 6)	68096	['activation_2[0][0]']
batch_normalization_2	(BatchNormalization)	(None, 112, 112, 25, 6)	1024	['separable_conv2d_1[0][0]']
max_pooling2d	(MaxPooling2D)	(None, 56, 56, 256)	0	['batch_normalization_2[0][0]']
conv2d_1	(Conv2D)	(None, 56, 56, 256)	33024	['activation[0][0]']
add	(Add)	(None, 56, 56, 256)	0	['max_pooling2d[0][0]', 'conv2d_1[0][0]']
activation_3	(Activation)	(None, 56, 56, 256)	0	['add[0][0]']
separable_conv2d_2	(SeparableConv2D)	(None, 56, 56, 512)	133888	['activation_3[0][0]']
batch_normalization_3	(BatchNormalization)	(None, 56, 56, 512)	2048	['separable_conv2d_2[0][0]']
activation_4	(Activation)	(None, 56, 56, 512)	0	['batch_normalization_3[0][0]']
separable_conv2d_3	(SeparableConv2D)	(None, 56, 56, 512)	267264	['activation_4[0][0]']
batch_normalization_4	(BatchNormalization)	(None, 56, 56, 512)	2048	['separable_conv2d_3[0][0]']
max_pooling2d_1	(MaxPooling2D)	(None, 28, 28, 512)	0	['batch_normalization_4[0][0]']
conv2d_2	(Conv2D)	(None, 28, 28, 512)	131584	['add[0][0]']
add_1	(Add)	(None, 28, 28, 512)	0	['max_pooling2d_1[0][0]', 'conv2d_2[0][0]']
activation_5	(Activation)	(None, 28, 28, 512)	0	['add_1[0][0]']
separable_conv2d_4	(SeparableConv2D)	(None, 28, 28, 728)	378072	['activation_5[0][0]']
batch_normalization_5	(BatchNormalization)	(None, 28, 28, 728)	2912	['separable_conv2d_4[0][0]']
activation_6	(Activation)	(None, 28, 28, 728)	0	['batch_normalization_5[0][0]']
separable_conv2d_5	(SeparableConv2D)	(None, 28, 28, 728)	537264	['activation_6[0][0]']
batch_normalization_6	(BatchNormalization)	(None, 28, 28, 728)	2912	['separable_conv2d_5[0][0]']
max_pooling2d_2	(MaxPooling2D)	(None, 14, 14, 728)	0	['batch_normalization_6[0][0]']
conv2d_3	(Conv2D)	(None, 14, 14, 728)	373464	['add_1[0][0]']
add_2	(Add)	(None, 14, 14, 728)	0	['max_pooling2d_2[0][0]', 'conv2d_3[0][0]']
separable_conv2d_6	(SeparableConv2D)	(None, 14, 14, 1024)	753048	['add_2[0][0]']
batch_normalization_7	(BatchNormalization)	(None, 14, 14, 1024)	4096	['separable_conv2d_6[0][0]']
activation_7	(Activation)	(None, 14, 14, 1024)	0	['batch_normalization_7[0][0]']
global_average_pooling2d	(GlobalAveragePooling2D)	(None, 1024)	0	['activation_7[0][0]']
dropout	(Dropout)	(None, 1024)	0	['activation_7[0][0]']
dropout	(Dropout)	(None, 1024)	0	['global_average_pooling2d[0][0]']
dense	(Dense)	(None, 1)	1025	['dropout[0][0]']

Table 4: Complete table of Convolutional Neural Network (CNN) architecture.

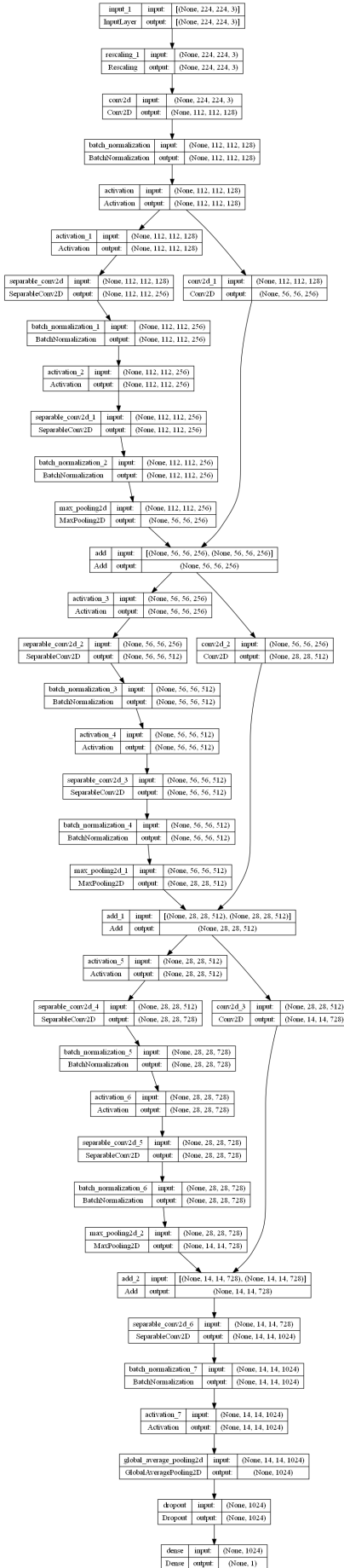


Figure 17: Diagram of Convolutional Neural Network (CNN) architecture.

### 10.3 Results of Basic Pathology Annotation

In this section we present the basic pathology annotations which we arrived at after having analyzed the images in Figure 16 (image file names are presented in Table 3).

#### 1. Image 1 – Basic Pathology Annotation 1 (BPA1)

(a) *Tumor Infiltrating Lymfocytes*

It is normal to have some Infiltrating Lymfocytes (ILs) in the epithelium, although not in too great of a quantity. There are many bacteria and other irritants in the gastric intestinal tract so it is special grounds for the training of the immune system. *Image 1* seems to contain some immune cells, but not too many. Possible ILs are annotated in red with an 'A' next to them. The low quantity of ILs is indicative of the image belonging to the MSS-class.

(b) *Chron's like reaction*

(c) *Poor differentiation*

Does all of the cell look like surrounding cells? [Yes, the cells and glands appear similar to each other in shape and colour.] The height of the cell: cylinder, cuboids, slice etc? [Yes, the cells and glands appear to be of normal height.] Nucleus shape and colour? [The resolution of the image is low, but from what is discernible there seems to be no signs of nuclear atypia or pleomorphism.] Chromatin packing or density? [The resolution of the image is low, but from what is discernible there seems to be no signs of abnormal chromatin packing.] Nucleus position in cell and in relation to other cells? [Different cells have different nucleus position, so what is normal for one type of cell can be abnormal for another. This makes our image difficult to interpret. The resolution of the image is low, but from what is discernible all visible nuclei seem to exhibit paracentral or periphic nuclei position. Possible nuclei are annotated in yellow with a 'C' next to them. This could indicate that the cells are secreting cells, cells that one could find in both colon and gastric tissues.] Quantity of mitosis? [The resolution of the image is low and no mitosis is discernible.] All things considered, *Image 1* appears rather normal, thus indicating that it belongs to the MSS class.

(d) *Mucinous or signet ring cell morphology*

Mucinous or signet ring cell morphology is only visible in very aggressive types and stages of cancer. *Image 1* seem to exhibit no signs of mucinous or signet ring cell morphology, this is thus indicative of the image belonging to the MSS-class.

## 2. Image 2 – Basic Pathology Annotation 2 (BPA2):

### (a) *Tumor Infiltrating Lymphocytes*

*Image 2* seem to contain some immune cells, but not too many. The possible ILs in the image are annotated red with an 'A' next to them. The low quantity of ILs is indicative of the image belonging to the MSS-class.

### (b) *Chron's like reaction*

### (c) *Poor differentiation*

Does all of the cell look like surrounding cells? [No, the cells and glands appear somewhat different to each other in shape and colour. An example of this is annotated yellow with a 'C1' next to it. That said, a general coherent structure still remains.] The height of the cell: cylinder, cuboids, slice etc? [No, the cells and glands appear to differ somewhat in height.] Nucleus shape and colour? [The resolution of the image is low, but from what is discernible there seems to be no signs of nuclear atypia or pleomorphism.] Chromatin packing or density? [The resolution of the image is low, but from what is discernible there seems to be no signs of abnormal chromatin packing.] Nucleus position in cell and in relation to other cells? [Different cells have different nucleus position, so what is normal for one type of cell can be abnormal for another. This makes our image difficult to interpret. The resolution of the image is low, but from what is discernible all visible nuclei seems to exhibit central nuclei position. Examples of these are annotated yellow with 'C2' next to them. This could indicate that the cells are muscle cells or epithelial cells, cells that one could find in both colon and gastric tissues.] Quantity of mitosis? [The resolution of the image is low, only one possible mitosis is discernible.] All things considered, *Image 2* appears to be more or less normal, thus indicating that it might belong to the MSS class.

### (d) *Mucinous or signet ring cell morphology*

Mucinous or signet ring cell morphology is only visible in very aggressive types and stages of cancer. *Image 2* seem to exhibit no signs of mucinous or signet ring cell morphology, this is thus indicative of the image belonging to the MSS-class.

### 3. Image 3 – Basic Pathology Annotation 3 (BPA3):

(a) *Tumor Infiltrating Lymphocytes*

*Image 3* seem to contain some immune cells, but not too many. These are annotated with red and with an 'A' next to them. The low quantity of ILs is indicative of the image belonging to the MSS-class.

(b) *Chron's like reaction*

(c) *Poor differentiation*

Does all of the cell look like surrounding cells? [Yes, the cells and glands appear more or less similar to each other in shape and colour. An example of this is annotated yellow with a 'C1/C2' next to it.] The height of the cell: cylinder, cuboids, slice etc? [No, the cells and glands appear to differ somewhat in height. An example of this is annotated yellow with a 'C1/C2' next to it.] Nucleus shape and colour? [The resolution of the image is low, but from what is discernible there seems to be no signs of nuclear atypia or pleomorphism.] Chromatin packing or density? [The resolution of the image is low, but from what is discernible there seems to be no signs of abnormal chromatin packing.] Nucleus position in cell and in relation to other cells? [Different cells have different nucleus position, so what is normal for one type of cell can be abnormal for another. This makes our image difficult to interpret. The resolution of the image is low, but from what is discernible all visible nuclei seems to exhibit central nuclei position. An example of this is annotated yellow with a 'C3' next to it. This could indicate that the cells are muscle cells or epithelial cells, cells that one could find in both colon and gastric tissues.] Quantity of mitosis? [The resolution of the image is low, no mitosis are discernible.] All things considered, 'Normal Image 1' appears rather normal, thus indicating that it belongs to the MSS class.

(d) *Mucinous or signet ring cell morphology*

Mucinous or signet ring cell morphology is only visible in very aggressive types and stages of cancer. *Image 3* seem to exhibit no signs of mucinous or signet ring cell morphology, this is thus indicative of the image belonging to the MSS-class.

### 1. Image 4 – Basic Pathology Annotation 4 (BPA4):

(a) *Tumor Infiltrating Lymphocytes*

*Image 4* seem to contain a few more immune cells compared to what we saw in the MSS-class images (*Image 1-3*). These are annotated in red with an 'A' next to them. The relatively high quantity of TILs could be indicative of the image belonging to the MSIMUT-class.

(b) *Chron's like reaction*

(c) *Poor differentiation*

Does all of the cell look like surrounding cells? [No, the cells and glands appear different in shape and colour.] The height of the cell: cylinder, cuboids, slice etc? [No, the cells and glands appear to differ in height.] Nucleus shape and colour? [The resolution of the image is low, but from what is discernible there seems to be different sized nuclei, indicative of nuclear atypia or pleomorphism. These are annotated in yellow with a 'C1' next to them.] Chromatin packing or density? [The resolution of the image is low, but from what is discernible there seems to be no signs of abnormal chromatin packing.] Nucleus position in cell and in relation to other cells? [Different cells have different nucleus position, so what is normal for one type of cell can be abnormal for another. This makes our image difficult to interpret. The resolution of the image is low, but from what is discernible the visible nuclei seems to have somewhat differently positioned nuclei. Compare nuclei annotated yellow and with a 'C1' next to them with those annotated with a 'C2' next to them.] Quantity of mitosis? [The resolution of the image is low, only two possible mitosis is discernible. See top yellow 'C1' ring annotation and bottom yellow 'C3' annotation.] All things considered, *Image 4* appears to have a rather high differentiation, thus indicating that it belongs to the 'MSIMUT' class.

(d) *Mucinous or signet ring cell morphology*

Mucinous or signet ring cell morphology is only visible in very aggressive types and stages of cancer. *Image 4* seem to possibly exhibit some signs of mucinous or signet ring cell morphology, although it is somewhat difficult to tell due to the low resolution and image patch scope – regardless of this, the possible mucinous or signet ring cell morphology is indicative of the image belonging to the 'MSIMUT' class. See orange annotated with 'D' next to them for examples.

## 2. Image 5 – Basic Pathology Annotation 5 (BPA5):

(a) *Tumor Infiltrating Lymphocytes*

*Image 5* seem to contain many immune cells compared to what we saw in the MSS-class images (*Image 1-3*). See areas annotated in red with an 'A' next to them for examples. The relatively high quantity of TILs could be indicative of the image belonging to the MSIMUT-class.

(b) *Chron's like reaction*

(c) *Poor differentiation*

Does all of the cell look like surrounding cells? [No, the cells and glands appear different in shape and colour.] The height of the cell: cylinder, cuboids, slice etc? [No, the cells and glands appear to differ in height.] Nucleus shape and colour? [The resolution of the image is low, but from what is discernible it is possible that there are different sized nuclei, indicative of nuclear atypia or pleomorphism. See areas annotated in yellow with 'C1' next to them for examples.] Chromatin packing or density? [The resolution of the image is too low to tell.] Nucleus position in cell and in relation to other cells? [Different cells have different nuclei position, so what is normal for one type of cell can be abnormal for another. This makes our image difficult to interpret. The resolution of the image is too low to discern anything.] Quantity of mitosis? [The resolution of the image is too low to discern any possible mitosis.] All things considered, *Image 5* appears to have a rather high differentiation, thus indicating that it belongs to the 'MSIMUT' class.

(d) *Mucinous or signet ring cell morphology*

Mucinous or signet ring cell morphology is only visible in very aggressive types and stages of cancer. *Image 5* seems to possibly exhibit some signs of mucinous or signet ring cell morphology, although it is somewhat difficult to tell due to the low resolution and image patch scope – regardless of this, the possible mucinous or signet ring cell morphology is indicative of the image belonging to the 'MSIMUT' class. Examples of these are annotated in orange with a 'D' next to them.

### 3. Image 6 – Basic Pathology Annotation 6 (BPA6):

(a) *Tumor Infiltrating Lymphocytes*

*Image 6* – in its interesting Manhattan like appearance – seem to contain relatively more immune cells compared to what we saw in the 'MSS' class images (*Image 1-3*). These are annotated in red with an 'A' next to them. The relatively high quantity of TILs could be indicative of the image belonging to the 'MSIMUT'

class.

(b) *Chron's like reaction*

(c) *Poor differentiation*

Does all of the cell look like surrounding cells? [No, the cells and glands appear somewhat different in shape and colour.] The height of the cell: cylinder, cuboids, slice etc? [No, the cells and glands appear to differ somewhat in height.] Nucleus shape and colour? [The resolution of the image is low, but from what is discernible there seems to be different sized nuclei, indicative of nuclear atypia or pleomorphism. These are annotated in yellow with a 'C1' next to them.] Chromatin packing or density? [The resolution of the image is too low to tell.] Nucleus position in cell and in relation to other cells? [Different cells have different nuclei position, so what is normal for one type of cell can be abnormal for another. This makes our image difficult to interpret. The resolution is low, but the position of the nuclei in the cells seem to differ somewhat from each other, although most are central.] Quantity of mitosis? [The resolution of the image is too low to discern any possible mitosis but two. These ones are annotated in yellow with a 'C2' next to them.] All things considered, *Image 6* appears to have a rather high differentiation, thus indicating that it belongs to the 'MSIMUT' class.

(d) *Mucinous or signet ring cell morphology*

Mucinous or signet ring cell morphology is only visible in very aggressive types and stages of cancer. *Image 6* seem to exhibit no signs of mucinous or signet ring cell morphology, this is thus indicative of the image belonging to the 'MSS' class.

## 10.4 Detailed results of comparison between Basic Pathology Annotation and eXplainable AI

In this section we present our notes on the differences between our basic pathology annotation and our XAI heat maps, as we noted them during visual comparison.

### 1. Image 1 – (BPA1) vs (GB1, VG1, GC1, SA1):

#### (a) *Guided Backpropagation*

*Guided Backprop. 1* (GB1) differed from *Basic Pat. Annot. 1* (BPA1) by not marking two of the presumed ILs, and was similar by marking the third lowest located IL. GB1 also put emphasis on the white inside of the glandular architecture, something which BPA1 ignored except for when judging the overall architecture and low differentiation of the image. GB1 was similar in the way it also marked all 'Cs' present in BPA1. Lastly, GB1 marked two interesting darker spots in the middle of the image, somewhat resembling an ongoing mitosis.

#### (b) *VarGrad*

*VarGrad 1* (VG1) somewhat continued the pattern of GB1, but being less spread out and emphasizing the whites inside of the glandular architecture.

#### (c) *Grad-CAM*

*Grad-CAM 1* (GC1) interestingly put a lot of emphasis on the glands themselves and how they curved when forming the larger coherent architecture of the tissue present in the image. This is something akin to looking for features indicative of low differentiation in BPA1. GC1 interestingly focused on the area to the top left, an area that BPA1 largely ignored. It is difficult to judge why the CNN found this area important. Perhaps due to the somewhat abnormally elongated gland located in the center of the "hottest" part of the heat map?

#### (d) *Sobol Attribution*

*Sobol Attr. 1* (SA1) focused on a possible IL at the top middle of the image. These is a feature that BPA1 also would have found important. What is less interpretable is the hot spot at the bottom of SA1. Perhaps the CNN found the somewhat jagged white area of some importance since it somewhat resembles signet ring cell carcinoma?

### 2. Image 2 – (BPA2) vs (GB2, VG2, GC2, SA2):

(a) *Guided Backpropagation*

*Guided Backprop. 2* (GB2) differed from *Basic Pat. Annot. 2* (BPA2) by not emphasising the presumed ILs marked in red with an 'A' in the bottom left corner of the image. Instead it spotted possible ILs which were not spotted in BPA2 in the epithelial glandular tissue in the middle and bottom right of the picture. GB2 did however find the areas marked in yellow with a 'C1' and 'C2' next to them in BPA2 somewhat important – perhaps indicating that features of visible possible nuclei as important for classification. One last interesting note is BPA2s emphasis on the somewhat odd "inlet" at the top left of the image.

(b) *VarGrad*

*VarGrad 2* (VG2) somewhat continued the pattern of GB2, but being less spread out and emphasizing the epithelial glandular tissue and the size colour and shape of its gland in two foci in the top and bottom left of the image. VG2 also interestingly emphasised the visible nuclei annotated as the lowest positioned 'C2' in BPA2.

(c) *Grad-CAM*

*Grad-CAM 2* (GC2) interestingly put a lot of emphasis on the glands themselves and how they curved when forming the larger coherent architecture of the tissue present in the image. This is something akin to looking for features indicative of low differentiation in BPA2. GC2 interestingly focused on the area containing a possible LI in the middle of the image.

(d) *Sobol Attribution*

*Sobol Attr. 2* (SA2) interestingly only focused on the bright pink cells below the epithelial glandular layer of the tissue, at the middle right part of the image. This was an area that was largely ignored in BPA2, except for when judging the images overall architecture and relatively low differentiation. These cells seem to be erythrocytes – red blood cells – indicating that this might be a section of a blood vessel, possibly a vein or artery.

### 3. Image 3 – (BPA3) vs (GB3, VG3, GC3, SA3):

(a) *Guided Backpropagation*

*Guided Backprop. 3* (GB3) differed from *Basic Pat. Annot. 3* (BPA3) by ignoring the possible IL annotated red with an 'A' next to it at the right side of the image. GB3 also seemed to ignore BPA3s top 'C3' annotations, instead focusing of the

area in-between the two yellow annotations. The bottom 'C3' was also somewhat de-prioritized, with the area above the bottom 'C3' prioritized instead. Lastly, 'C1/C2' failed to capture much of GB3s areas of importance. It seems like GB3 focused a lot on sudden changes in the glandular architecture and on differentiation as a whole. GB3 also succeeded in spotting some potential ILs in the top left corner which were not annotated, although mentioned as important features, in BPA3. Both GB3 and BPA3 found the possible mitosis in the top 'A' as feature worth noting.

(b) *VarGrad*

*VarGrad 3* (VG3) somewhat continued the pattern of GB3, but being less spread out and emphasizing the bottom part of the image instead of the top like in GB3. VG3 especially focused on the area in 'C1/C2' in BPA3 in-between the two glandular formed "walls" in a brighter colour. It is difficult to say what makes this area important for the CNN. Perhaps it has to do with the general regularity of the area, exposing signs of low-differentiation associated with class prediction in BPA3?

(c) *Grad-CAM*

*Grad-CAM 3* (GC3) seems to try to establish the overall structure of the sample present in the image, but seems to struggle somewhat to do so, perhaps indicating signs of higher differentiation than what was proposed in BPA3. A sign that this hypotheses might be reasonable is the heat map quarter circle following the quarter circle of glands in the bottom left corner.

(d) *Sobol Attribution*

*Sobol Attr. 3* (SA3) interestingly focuses on the bottom 'C3' area as annotated in BPA3. Other than that, only the large area of 'C1/C2' capture any of SA3s areas of importance. SA3 also recognizes a visible possible nuclei in the middle of the image missed in BPA3, but a feature that BPA3 found important.

1. **Image 4 – (BPA4) vs (GB4, VG4, GC4, SA4):**

(a) *Guided Backpropagation*

*Guided Backprop. 4* (GB4) differed from *Basic Pat. Annot. 4* (BPA4) by ignoring most of the BPA4 possible ILs, annotated in red with an 'A' next to them, and

instead highlighting other potential ILs which BPA4 missed. GB4 also seemed to ignore most of BPA4s yellow 'C1' and 'C2' annotations. What BPA4 did notice however, was 'C3' and its interesting chromatin packing or nuclei split. GB4 also noticed all of BPA4s 'D' annotations, yet put more emphasis on other examples of the feature than in BPA4. GB4 put especially large focus on the signet ring cells in the top right and bottom left of the image.

(b) *VarGrad*

*VarGrad 4* (VG4) differed from (GB4) by focusing on larger details/features in the image than in VG4 which tended to trace the walls of each signet ring cell. VG4 focused much more in the middle than GB4. From top to down, VG4 focused on a interesting splitting signet cell in the top right. On a faded out, hard to spot, signet cell just above 'C2'. Another faded signet ring cell to the upper left of the bottom 'D'.

(c) *Grad-CAM*

*Grad-CAM 4* (GC4) differed by focusing on larger features still, and only highlighting areas of importance instead of any details. GC4 is difficult to interpret, but it seems that it is looking for ILs and glands.

(d) *Sobol Attribution*

*Sobol Attr. 4* (SA4) was similar to GC4 in that it focused on larger areas rather than detailed features. SA4 seems to have captured a more interpretable pattern as compared to the "pattern" produced by GC4. At closer inspection, even this SA4 pattern is difficult to interpret, although it seems to focus on areas with relatively low "density" of contrast.

## 2. Image 5 – (BPA5) vs (GB5, VG5, GC5, SA5):

(a) *Guided Backpropagation*

*Guided Backprop. 5* (GB5) seemed to agree with most of the areas in *Basic Pat. Annot. 5* (BPA5) – except for the top 'A' and the 'C1's. GB5 also shows how the CNN finds more blurred and bled out signet ring cells to be of particular importance, take 'D' and the "ribbon" of signet ring cell to the top right of the image as two examples. Except for this, GB5 also seemed to look for possible mitosis and ILs as seen in the area above the top 'C1', the area below the top 'D' and in the bottom left of the image.

(b) *VarGrad*

*VarGrad 5* (VG5) focused mostly on blurred, cloudy segments of signet ring cells.

(c) *Grad-CAM*

*Grad-CAM 5* (GC5) is difficult to interpret but seems to look for ILs.

(d) *Sobol Attribution*

*Sobol Attr. 5* (SA5) is difficult to interpret, but seems to look at a lot of different things – the most highlighted one being a patch of high contrast material in the muscle tissue between the image’s two main glandular areas.

**3. Image 6 – (BPA6) vs (GB6, VG6, GC6, SA6):**

(a) *Guided Backpropagation*

*Guided Backprop. 6* (GB6) seemed to agree with most of the areas in *Basic Pat. Annot. 6* (BPA6). It seems to be searching for high contrast areas indicative of ILs and cell nuclei.

(b) *VarGrad*

*VarGrad 6* (VG6) is difficult to interpret. It seems to possibly be somewhat confused by the large amount of white areas in the image, possibly miss-interpreting parts of this as white ring cells.

(c) *Grad-CAM*

*Grad-CAM 6* (GC6) continues to be, like VG6, difficult to interpret. GC6, like VG6, focuses on the white areas of the image – possibly confusing these areas for white ring cells. It particularly seems to mark ”wrinkles” and ”grooves” in the whitish tissue.

(d) *Sobol Attribution*

*Sobol Attr. 6* (SA6) is difficult to interpret. SA6 marks bigger areas rather than detailed features.