Explainable AI in Pathology: Concept Based Explainability for Mitotic Figure Detection in Whole Slide Images

CA4021 BSc Data Science - Formal Project Proposal

Adam Tegart

19327493

adam.tegart2@mail.dcu.ie

Supervisor: Alessandra Mileo

Abstract

Artificial Intelligence (AI) is quickly becoming a ubiquitous entity in society with unfathomable accuracy across a range of applications. There is plenty of good that can be done in the area of healthcare in particular, but in this highly regulated space these models have to be understood. In many applications there is little interest in converting the black box model to a glass box that is transparent and easily interpretable. For health applications, every necessary action must be taken to ensure patient safety. Making use of models that cannot be understood is not an ethically acceptable action when working with individuals' livelihoods. Finding a remedy to this issue is the only way to promote trust and acceptance of AI in these regulated use cases and ensure that the potential of AI for good is not left untapped. This paper explores the use of human-interpretable concepts to explain the predictions made by object detection models trained to detect mitotic figures. We will create a visual representation of these concepts within the input image to ensure that the model understands the concept as we do. We will also look at how relevant these concepts are for a model trained on a different dataset in the same domain. The primary goal is to better evaluate the quality of these explanations.

Section 1: Motivation and Background

AI has had a very impactful presence in our lives in recent years, with the increase in computing power we can now make better use of AI than ever before. This increase in compute power has led to neural networks that are drastically bigger both in size and complexity, which in turn leads to networks that can more accurately model a problem with an adequately large amount of data. OpenAI have made waves recently with the release of a conversational model built on their GPT-3 language model, ChatGPT [1]. This conversational model can be seen to produce very human-like responses and highlights the advancements in AI made in areas of human intelligence, such as linguistics. Autonomous vehicles are another area of interest as advancements have been made in the area of computer vision [2]. As computer vision networks become more sophisticated and the data available grows, we can see an increase in their capabilities to approach these tasks that a decade ago seemed infeasible. These milestones that have been achieved can be attributed to the increase in compute power and availability of data that allowed neural networks to be harnessed to the full extent of their potential. These advancements have made using AI very desirable and this will continue to be the case if current trends continue.

The primary drawback of the increase in accuracy due to a more complex model is that the predictions made by the model become difficult to interpret [3, 4]. Models of this nature are considered a black box as the model is given input and returns a prediction. The steps taken to produce the output are not easily explainable, leaving the inner workings of the model obscured. Transparency in AI models is an important concept when these models are placed within important infrastructure in our society. Explainability within models can allow us to better understand how they operate and use this as feedback to improve performance. It can also be used to improve our confidence in predictions when we intend to act on the insight in situations that impinge on a person's autonomy or impact their livelihood in any way. It is for this reason that eXplainable AI (XAI) has become an area of focus for many researchers. It can be seen that there has been an exponential increase in papers published in the area of XAI since 2016 [5], which is a testament to the importance of explainability for the future of AI and the continuation of research in the area. Making headway in eliminating this black box model will allow for transparency and promote more trust in AI when using insight for decision making. In addition, it will likely deepen our understanding as the rationale behind insights can be explained and added to our own understanding of a problem.

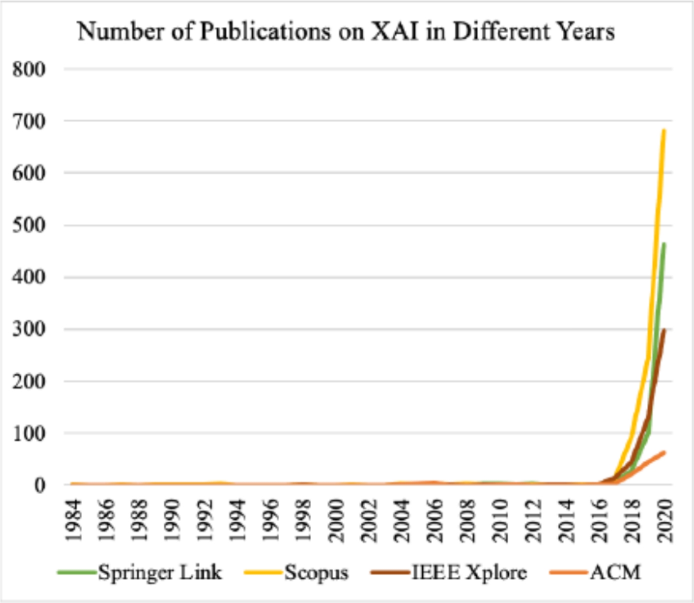

Figure 1.1. Number of publications from several renowned bibliographical databases, via Islam, M.R et al. 2022 [5].

In the field of healthcare, AI could have immense benefits for patients if implemented safely and with care. Ensuring that the output from models can be explained is at the heart of this safe implementation. There is no place for these models in healthcare when the decisions have weight on lives and lack interpretability. According to Statistica, “In 2021, 42 percent of healthcare organizations in the European Union were currently using AI technologies for disease diagnosis, while a further 19 percent had plans to employ this technology within the next three years” [6]. So adequate XAI methods must be in place to ensure that the decisions made that impact an individual's livelihood can be explained to them as a healthcare professional would.

According to PubMed [7], there is still an increase in the imaging examinations carried out year after year, but there has been a decrease in pace. This showcases the need for AI in healthcare to alleviate the workloads of healthcare staff at a time when the workload is increasing due to the general population living longer. There is a need for the prediction from an AI model to be explainable for the sake of transparency. We can use this explanation to justify or dismiss the prediction by using our own judgement, or possibly even further our own knowledge of a problem. This highlights why robust and fair XAI methods are needed to ensure the use of AI in healthcare is not a risk to the wellbeing of patients and is sufficiently transparent.

Section 2: The Problem Statement

This project will look at how explainability can be applied to the task of detecting mitotic figures within histopathology slides. Mitotic figures are cells that are undergoing mitosis and in large numbers can be indicative of a cancerous region of tissue. A pathologist typically considers many visual features of a cell to determine whether it is undergoing mitosis, and there are several stages within the cycle that have different appearances [8]. Manually inspecting cells to determine a count of mitotic figures within a region can be an arduous task. Humans are also prone to error and there is the possibility for inter-observer variability between pathologists. This motivates the use of AI to generate a count of the mitotic figures within a region, but this has to be done with transparency in mind. A pathologist must have trust in the model to correctly find the mitotic figures, and it is upon explainability that this trust is built. The use of explainability is important as a pathologist can validate that the model is considering the correct features when detecting mitotic figures, ensuring that the model is more accurate as the predictions are based on standard practice.

The objective for this project is to extract and validate a set of human observable concepts that are used to classify mitotic figures from the Mitosis Domain Generalisation Challenge [9] dataset. After training an adequate model on the data the impact of these concepts on specific predictions will allow us to determine how impactful they are when detecting a mitotic figure. Then these concepts will be taken and tested against a separate dataset to assess how generalizable they are across different datasets. We also plan to make use of a visualisation to help convey the impact a concept has based on the Gifsplanation implementation [10]. The project itself will look at a specific use case, but the parts it consists of can be applied to any image based task in theory.

This project will help to investigate how effective these extracted concepts are when interpreting the predictions and also how well they generalise across different datasets. This will allow for better feedback on examples the model can improve on as we can understand how to model is computing a prediction. In addition, we can also quantify the confidence we have in our model to predict the correct class based on the correct criteria. This is to say, we can be confident that a model detects and classifies a mitotic figure using the same visual attributes that a pathologist would. This approach will lead to more robust models that we can be confident are accurate for all the right reasons.

Section 3: State of the Art

There are two areas that are relevant for contextualising the contribution of this project, one more important than the other. Firstly, we must decide the best architecture of the neural network for the task of object detection. This is not the primary focus of the project, so we don't need to expend too many resources here as we are not too concerned with the performance of the model, but it is still worth understanding the state of the art. The primary focus will be on the explainability techniques and how these can be applied to tasks involving images, mainly in the field of medical imaging. This will help to guide the efforts of this project and ensure that the most meaningful results can be achieved for the defined problem.

Object Detection and Modern Architectures for Models

Object detection has been a very well researched computer vision task over the last 20 years [11, 12]. The basic premise is to take a step beyond classification of objects within an image and predict the location of objects within an image. This is typically done by returning the coordinates of a box that encloses the object along with the class of the object. For this reason, it is a more complex task to approach than classification of an image based on classes.

Pioneers

Initially, the task was approached in a way that involved a heavy focus on feature engineering. This was due to limitations with regard to compute power at the time in the 2000's. This resulted in methods such as Histogram of Oriented Gradients [13] and Deformable Part-based Model [14], which made efficient use of the resources available to detect objects. HOG acts as a feature extractor and creates a representation of the image that contains only the angles and their magnitudes, this acts as a form of edge detection. DPM is an approach that expands on HOG by using a "divide and conquer" approach. It consists of a root-filter and several part-filters whose size and location are configured by a weakly-supervised model. Although these approaches are not the current state of the art, they have been highly influential in the field.

Convolutional Neural Networks

The next advancements in object detection came in the form of convolutional neural networks. These allow for the model to learn how to extract robust features for the specific task, eliminating the need for hand-crafted features that are based on our understanding. R-CNN [15] was the first object detection model proposed to tackle object detection. The premise is that 2,000 regions are selected using a selective search algorithm that joins similar regions together in a greedy fashion. These regions are re-sized and passed to the CNN to extract the features of interest and the bounding box coordinates. The features are passed to a SVM to determine if an object is present in that region. This whole process is computationally expensive and cannot be computed in real-time as a result. SPPNet [16] made improvements on R-CNN, primarily by introducing a spatial pyramid pooling layer. This allows for the feature map to be computed once for the image and then representations for arbitrary regions to be extracted from this complete feature map. This significantly improves the training time for R-CNN.

Fast R-CNN [17] takes the improvements of SPPNet and speeds up the training and inference times. The SVM is replaced with a softmax layer to predict the object class and return a bounding box. The bottleneck is now at the region proposal stage, the selective search algorithm is fixed and does not learn. Faster R-CNN [18] focuses on improving the region proposal stage by using a region proposal network. This network is trained like any other and allows for low-cost region proposals which significantly speeds up Fast R-CNN, making it near real-time to inference. Feature pyramid networks (FPN)[19] have been introduced as a way to incorporate semantic information from several layers of the network into the prediction. This is carried out by using lateral connections between the different layers in the feedforward pass of the network and the layers created by upsampling the coarse feature maps produced. This adds more high-level semantic information to the prediction at the cost of a substantial increase in computation.

The models mentioned up to this point have been two-stage approaches, the regions are proposed and then the objects are detected within these regions. There has been much work in recent years with regard to one-stage detectors, which do not require image proposals. The pioneer for this is YOLO [20], which divides the full image into an S x S grid. Every cell in this grid was responsible for detecting objects whose centre resided in that cell. A class probability is returned for each cell and these are used to determine the location of objects within the image. The main limitation of YOLO is that it struggles to localise smaller objects accurately. There have been improvements to YOLO, with v2, v3 and v4 [21, 22, 23] being introduced. These improvements included changing the feature extractor to Darknet-19 and then Darknet-53, batch normalisation and data augmentation. The most impactful changes are from YOLOv4, which introduces “bag of freebies” and “bag of specials” methods. The “bag of freebies” methods only increase the training cost for an increase in accuracy, these include data augmentation, self-adversarial training and class label smoothing to name a few. The “bag of specials” methods add computation to the inference for an increased accuracy, these include Mish activation [24] and SPP-Block [25] to name a couple.

Looking quickly at the other single-stage models, Single Shot Multibox Detector [26] was the first model to use different layers in the network to detect objects of different scales. RetinaNet [27] introduced “focal loss” as a new loss function to put more focus on hard misclassified examples, which helped to alleviate some of the issues caused by foreground-background class imbalance. CentreNet [28] makes use of a stacked hourglass-101 [29] backbone and predicts the centre of an object, estimates the size of the object and then the offset of the object point is corrected. EfficientDet [30] improves upon FPN with BiFPN, a bi-directional implementation of FPN.

Backbones

The backbone, also known as the feature extractor, is an integral part of an object detection model. AlexNet [31] reintroduced convolutional neural networks. VGG [32] focused on the depth of the architecture to increase the receptive field. Inception [33] was introduced to tackle the issue of kernel size, when an object fills most of the image a large kernel is preferred, when an object is small in relation to the image size a small kernel would be preferred. The inception module, which the Inception architecture consists of, applies several different sized kernels and global pooling to the same input feature map and concatenates the output. This allows the model to work well for both small and large objects. ResNet [34] introduced skip connections to stacks of convolutional layers, this aims to reduce the decay of performance for deeper networks. ResNeXt [35] combines the ideas introduced in Inception and ResNet to create an architecture that uses the residual blocks from ResNet in a “split-transform-merge” fashion like Inception. EfficientNet [36] introduced uniform scaling for model width, depth and image size. The intuition being that a larger image needs more layers for a bigger receptive field and more channels to capture more patterns within the image.

Transformers have had a huge impact in the field of natural language processing, this has led to questioning if transformers can be applied to computer vision tasks. There has been progress in this space with ViT [37], DeTR[38] and Swim Transformer [39], though these require considerable data, have more parameters and lack inductive bias.

Explainable AI in medical images

Modern healthcare and advances in this area have led to better patient outcomes. AI has become an entity that spans across many sectors, and the healthcare sector is not immune. Non-invasive medical imaging techniques have become more popular and this has led to an increase in medical imaging data. Explainability is key when using AI in these circumstances where human lives hang in the balance. So, we will look at the different types of explainability techniques that can be seen in the literature for AI in the context of medical imaging [40, 41, 42].

Visual explanation

Visual explanation methods involve the use of saliency maps or masks that identify the parts of an image that played a role in the prediction. These can be created by using a backpropagation or perturbation approach.

Backpropagation Approaches

The earliest approaches made use of partial-derivatives to generate class images and image specific saliency maps [43], deconvolution steps to reverse the steps taken by a convolutional neural network [44] and guided-backpropagation [45] which makes use of both vanilla backpropagation and deconvolution. Guided-backpropagation acts as vanilla backpropagation except at ReLUs, where it only keeps positive error signals as in deconvolution.

Class activation maps (CAM) [46] alter the structure of a neural network and replace the fully connected layer with global pooling. Taking the weighted sum of the inputs given to the neuron that won reveals where the network was focused when making the prediction of that class for the given image. Grad-CAM [47] does not require any model architecture modifications unlike CAM. Grad-CAM makes use of the gradients to allow for the use of the fully connected layers. There is also the option to select a layer to use the feature maps from. Guided Grad-CAM [47], takes Grad-CAM and applies methods from guided-backpropagation to create a more fine-grained visualisation. These methods work well for understanding which features in an image activate the winning neuron in the softmax layer.

Layerwise relevance propagation (LRP) [48] acts similarly to CAM, but is model-agnostic as long as the model is structured as a neural network. The prediction is back propagated through the network until it reaches the input. Deep SHAP [49] makes use of SHapley Additive exPlanation (SHAP) values [50], these were first used in game theory to determine the contribution of each player by measuring performance with and without them included. This can be applied to CNN's by making use of DeepLIFT [51] to approximate the SHAPley values for the network's input, improving the performance.

Attention masks are another way of determining how a model makes a prediction [52]. This involves computing the compatibility of a patch within the image against the global image, this is done by comparing the patch's feature vector against the global feature vector. This can aid in understanding what influenced the decision and can be used to ensure the correct information was used to determine the prediction.

Perturbation Approaches

Perturbation approaches perturb the input slightly to understand the importance of different features in the input. Local interpretable model-agnostic explanations (LIME) [53] makes use of a simple model to approximate the complex model locally. This involves perturbing the input slightly and gathering the output from the complex model. A simple model is then trained on this set of perturbed inputs and corresponding outputs. In this way, the complex model can be approximated in this local region around the output.

Occlusion sensitivity [54] is a perturbation technique that occludes part of the image and assesses the quality of the prediction, in this way the pixels that have the most impact can be found. Meaningful perturbations [55] are as they sound, the image is altered in a plausible fashion by using a constant value, blur or noise. It has been found that this approach is not suited for medical imaging as these natural phenomena are typically present anyways, or implausible in the case of the constant value.

Generative adversarial networks (GANs) [56] have been used as a method to produce counterfactual examples. CycleGAN [57] is an approach that takes an image X in some domain and aims to translate it into Y in the target domain. CycleGAN is trained in a way that allows images to be translated from one domain to the other and vice versa, making it ideal for generating counterfactuals. Gifsplanation [10] generates counterfactuals in a similar way, the latent representation is moved in the most semantically meaningful direction for a classification. This can be seen as a GIF that visualises the counterfactual and highlights the changes necessary to switch the prediction.

Multiple Instance Learning

Multiple instance learning (MIL) is an approach that typically uses a bag of instances that is labelled. The task is to learn from these bags and be able to classify future bags based on these labelled bags already seen. In medical imaging, MIL has been adapted to use the whole image as a bag, and the patches as instances [58]. In this way, the whole image can be classified based on the patches in a supervised way that only needs the image to be labelled. This can be seen as explainable as we can see which instances in the bag are contributing to the prediction.

Latent Space Interpretation

The latent space is a compressed version of the features in the input. This is typically carried out to represent the features of the input in a more cost effective way and allow similarity between instances to be computed cheaper. Methods such as principal component analysis and t-distributed Stochastic Neighbour Embeddings (t-SNE) [59] allow for the dimensionality to be reduced with much of the information preserved. Autoencoders can be used to create such a latent representation of an input, allowing for interpretability if used with a trained model [60]. If the concepts used to classify the input are encoded, we can visualise these with a decoder to recreate the original image. We can assess this qualitatively by looking at the original image and the recreated image, or quantitatively by passing the recreated image through the original model.

Textual explanation

Textual explanations are inherently human-interpretable. Introducing a justification module onto the model [61] is a way of creating such a textual explanation. This justification module makes use of an intermediate feature map extracted from the model and the final prediction to generate both a text and visual explanation of the specific prediction.

Testing with Concept Activation Vectors (TCAV) [62] is an approach that measures the sensitivity of models to certain human-interpretable concepts. A concept activation vector is found by gathering examples with and without the concept and training a linear model. The normal to the hyperplane separating these 2 groups is considered the concept vector. This was improved with Regression Concept Vectors [63], which reframes the problem as a regression problem. Regression concept vectors are found by least squares linear regression on the concept measures. These are bidirectional unlike the CAVs, and allow for the use of continuous features to determine concepts. Automated Concept-based Explanation (ACE) [64] is a method to derive and cluster concepts from a set of images and determine their importance using TCAV. The benefit is that only the cluster would need to be labelled to understand the set of concepts produced.

Case-based explanation

Explaining a prediction using other examples is a very human approach to explainability. ProtoPNet [65] takes patches of an input image and computes the similarity with prototypical examples learned during training. These can then be given along with the prediction to indicate what parts of the input were similar to prototypes seen prior during training.

Evaluation of the explanations

This is an area that has yet to become standard practice in XAI papers and will likely become more relevant in the coming years. There has been a framework proposed for a qualitative analysis [66]. There are 3 forms of evaluation, application-grounded, human-grounded and functionality-grounded. Application-grounded involves experiments in the real use case with human experts involved, this would mean a medical professional for medical image analysis. Human-grounded only requires human assessment, they need no expertise in the subject matter. This allows for a less expensive evaluation that still assesses the general quality of explanations. Functionality-grounded evaluation uses other proxies to determine the quality, such as annotations of areas that are relevant to a prediction which can be compared against visual explanations.

Section 4: Methodology

My approach will aim to better understand how well concept-based explainability works for this particular use case where there can be large variability of small objects. This project also aims to understand how generalizable these concepts derived automatically are across datasets by assessing the relevance for a model trained on a public dataset and another trained on a proprietary dataset from Deciphex. These concepts will also be visualised using GIFs as a medium to understand what the changes of a concept signify for the model.

The data that will be used to train my object detection model is publically available. The MItosis DOmain Generalization Challenge 2022 dataset [9] consists of ~9,500 mitotic figure annotations and ~11,000 hard examples that are not mitotic figures. These are spread across several different tissue types such as lung, breast, skin and lymphatic for humans and canines. There is variability in the annotations due to the different colouration of the tissues which should help any model trained on the data more generalizable.

The model I plan to use will be a Faster R-CNN implementation with a ResNet50 backbone for feature extraction. This model implementation is available in the TensorFlow2 Detection Model Zoo on GitHub [66]. This implementation was chosen as accuracy is not the primary objective of this project and the general structure is similar to the proprietary model that I will have access to from Deciphex. This will act to reduce the number of factors that can affect the results and cause disagreement between the 2 models. The training of this model will be carried out on a server that I will be given access to by my supervisor. The model will be assessed using standard metrics for object detection tasks such as the AP for mitotic figures. This is necessary as we want to ensure that the model has learned how to detect a mitotic figure before we probe it to see if it is considering the correct concepts.

As I am training the model I will set up the infrastructure needed to conduct the experiments I would like to run. This will include setting up the Automated Concept-based Explanation [64] code, Gifsplanation [10] and Testing with Concept Activation Vectors [62]. These will be needed to derive the concepts and test how influential they are in predictions. I will be making use of Pycharm locally to edit code and I will use Google Colab if needed. I will be making use of the OpenSlide python library to open the .tiff files that the whole slide images are supplied in. Tensorflow will be the library I use to train my object detection model. I will likely make use of Pandas and Numpy at times when performing evaluation of both the model and the concept vectors.

I will use Automated Concept-based Explanation [64] to avoid having to manually label concepts relating to mitotic figures that I lack the expertise to carry out. These extracted concepts will be clustered into groups that I can have a contact within Deciphex assess and label on my behalf. These concepts can then be used as they are within the Testing with Concept Activation Vectors [62] paper. I expect the result to be that some of the more visually obvious concepts have a detectable influence on positive predictions. I will make use of the visual features of mitotic figures [8] to aid my evaluation of the explanations.

To accompany the explanations offered by the TCAV methods I will use a visual method. In the same way that the Gifsplanation [10] allows the counterfactual to be displayed, I would like to show the difference caused by the change in presence of the concept. This will involve altering the latent representation of the input image by some factor of the concept vector to create a new representation where the concept is more prevalent. This will allow for assessment of the concept against a prediction knowing how the visual would change with different levels of the concept present. This task will involve making use of GANs as Gifsplanation did to create a latent representation that can then be altered and decoded into a new image that represents the image with a different level of the concept.

Following on from the use of my own model, I would like to apply the concepts derived from the public MIDOG [9] dataset to the proprietary model from Deciphex. I would like to assess the relevance of the concepts derived on a model that has been trained on a different dataset in the same domain. There will be differences due to scanner variability, staining and different tissue types. I expect that the results will be that there is some relevance, but the datasets are not representative enough of the entire problem to coincide with one another. This will give useful insight into how transferable concepts are between datasets of different origin.

The research questions I would like to answer primarily involve the transferability of concepts between datasets. The quality of concepts has already been studied so this is not an area of interest, though I would like to see how well the concepts are for mitotic figures which can be quite complex and difficult to discern by eye. I would like to make use of both the Testing with Concept Activation Vectors [62] and the Regression Concept Vectors [63] to create concept vectors as both approaches vary. I will make use of my Gifsplanation variation to assess how the concepts change within the same prediction to assess which approach is more accurate. There is more scope for additional questions as I am sure I will encounter new ideas as I run these experiments.

The evaluation methods for the explanations will be qualitative in nature. I will review the explanations myself in a human-grounded fashion. This will provide some general assessment of the quality with regard to ensuring the predictions are reasonable. I may also get the opportunity to carry out an application-grounded evaluation if I can get contacts in Deciphex to review the explanations and apply their knowledge of the features of mitotic figures. There will be little time for any consideration of functionality-grounded evaluation; it would involve manual annotations to compare against and I lack the expertise to carry this out. Overall, I expect that this evaluation will cover most of the concerns relating to the quality of the explanations.

All of the planned tasks have been fit into a timeframe prior to the deadline, leaving 2 weeks to act as cushion time. This time may be needed if I run into any issues with the setup or carrying out experiments. I may also use this time for additional experiments that may not currently be in scope if I made good progress.

This concludes the methodology, as the project develops it is likely I may deviate from this defined plan as I will have to adapt to any changes to occur or new information that changes the best course of action. Though I will be making use of concept vectors and creating some form of visual representation of these concepts that mimic that of the Gifsplanation [10] implementation.

Section 5: Project Plan

Below is the Gantt chart for my project. The chart is broken down into significant milestones in the project and is spread across the week in 2023.

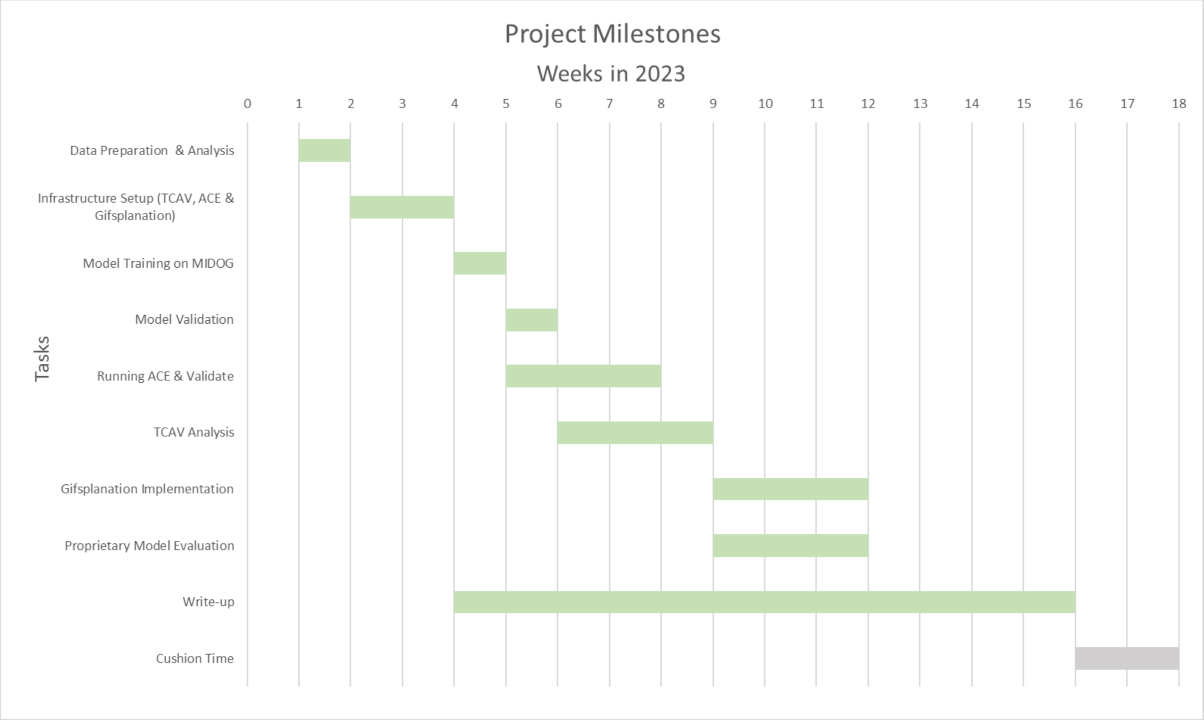

Figure 5.1 Gantt chart for this project

Data Preparation & Analysis - Ensure that data is in a usable format. Open the data and become familiar with it. Become familiar with the python library to open whole slide images, OpenSlide.

Infrastructure Setup (TCAV, ACE & Gifsplanation) - Find the code from the various papers I intend to make use of. Ensure that the code functions as expected and become familiar with it.

Model Training on MIDOG - Train the object detection model on the training data. This will be done on a server that Alessandra has access to.

Model Validation - Assess the accuracy of the model. Review some predictions and note the behaviours.

Running ACE & Validate - Run ACE to automatically extract concepts. Ensure that these seem reasonable and confirm the concept type with contact in Deciphex as this is not my field.

TCAV Analysis - Analyse the TCAV results from ACE and determine how relevant they are. Note how well they were applied and derive insight into the usefulness.

Gifsplanation Implementation - Create visualisations that display the concepts in a fashion similar to Gifsplanation.

Proprietary Model Evaluation - Evaluate the model that will be supplied by Deciphex. Determine if the concepts from the MIDOG dataset are applicable and if so to what degree? Can the predictions be adequately explained?

Write-up - Time to spend constructing my report, this will begin early and be completed as milestones are completed.

Cushion Time - An additional 2 weeks at the end of the project before May to account for any issues I may encounter and need more time for.

References

[1] OpenAI, “ChatGPT: Optimizing Language Models for Dialogue,” *OpenAI*, Nov. 30, 2022. https://openai.com/blog/chatgpt/ (accessed Jan. 02, 2023).

[2] L. Liu *et al.*, “Computing Systems for Autonomous Driving: State-of-the-Art and Challenges,” *arXiv.org*, 2020, doi: 10.48550/arXiv.2009.14349.

[3] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," in IEEE Access, vol. 6, pp. 52138-52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[4] A. Barredo Arrieta *et al.*, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[5] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, “A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks,” *Applied Sciences*, vol. 12, no. 3, p. 1353, Jan. 2022, doi: 10.3390/app12031353.

[6] “Adoption stage of AI in healthcare in the EU 2021 | Statista,” *Statista*, 2021. https://www.statista.com/statistics/1312566/adoption-stage-of-ai-in-healthcare-in-the-eu/ (accessed Jan. 02, 2023).

[7] R. Smith-Bindman *et al.*, “Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016,” *JAMA*, vol. 322, no. 9, p. 843, Sep. 2019, doi: 10.1001/jama.2019.11456.

[8] A. Lashen *et al.*, “Visual assessment of mitotic figures in breast cancer: a comparative study between light microscopy and whole slide images,” *Histopathology*, vol. 79, no. 6, pp. 913–925, Sep. 2021, doi: 10.1111/his.14543.

[9] Marc Aubreville, Christof Bertram, Katharina Breininger, Samir Jabari, Nikolas Stathonikosand Mitko Veta, “MItosis DOmain Generalization Challenge 2022”. Zenodo, Mar. 16, 2022. doi: 10.5281/zenodo.6362337

[10] J. P. Cohen *et al.*, “Gifsplanation via Latent Shift: A Simple Autoencoder Approach to Counterfactual Generation for Chest X-rays,” *arXiv.org*, 2021, doi: 10.48550/arXiv.2102.09475.

[11] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, “A survey of modern deep learning based object detection models,” *Digital Signal Processing*, vol. 126, p. 103514, Jun. 2022, doi: 10.1016/j.dsp.2022.103514.

[12] Z. Zou, Z. Shi, Y. Guo, and J. Ye, “Object Detection in 20 Years: A Survey,” *arXiv.org*, 2019, doi: 10.48550/arXiv.1905.05055.

[13] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1. IEEE, 2005, pp. 886–893.

[14] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008, pp. 1–8.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Region-based convolutional networks for accurate object detection and segmentation,” IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 1, pp. 142– 158, 2016

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual 29 recognition,” in European conference on computer vision. Springer, 2014, pp. 346–361.

[17] R. Girshick, “Fast r-cnn,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[18] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in Advances in neural information processing systems, 2015, pp. 91–99.

[19] T.-Y. Lin, P. Dollar, R. B. Girshick, K. He, B. Hariharan, ´ and S. J. Belongie, “Feature pyramid networks for object detection.” in CVPR, vol. 1, no. 2, 2017, p. 4.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[21] J. Redmon and A. Farhadi, “Yolo9000: better, faster, stronger,” arXiv preprint, 2017.

[22] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” arXiv preprint doi:1804.02767, 2018.

[23] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” *arXiv.org*, 2020, doi: 10.48550/arXiv.2004.10934.

[24] D. Misra, "Mish: A self regularized non-monotonic neural activation function." arXiv preprint, vol. 4, no. 2, 2019, doi:10.48550/arXiv.1908.08681.

[25] . He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916, https://doi.org/10.1109/TPAMI.2015.2389824.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in European conference on computer vision. Springer, 2016, pp. 21–37.

[27] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 318-327, 1 Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

[28] X. Zhou, D. Wang, and P. Krähenbühl, “Objects as Points,” *arXiv.org*, 2019, doi: 10.48550/arXiv.1904.07850.

[29] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose es-

timation, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, in: Lecture Notes in Computer Science, Springer International Publishing, 2016, pp. 483–499.

[30] M. Tan, R. Pang, Q.V. Le, EfficientDet: scalable and efficient object detection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 10778–10787, https://ieeexplore.ieee.org/document/9156454/.

[31] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2012, p. 9.

[32] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv.org*, 2014, doi: 10.48550/arXiv.1409.1556.

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, arXiv:1409.4842.

[34] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[35] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, arXiv:1611.05431.

[36] M. Tan, Q.V. Le, EfficientNet: rethinking model scaling for convolutional neural networks, arXiv:1905.11946.

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, arXiv:2010.11929, 2021.

[38] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, arXiv:2005.12872, 2020.

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, arXiv:2103. 14030, 2021.

[40] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, “Transparency of deep neural networks for medical image analysis: A review of interpretability methods,” *Computers in Biology and Medicine*, vol. 140, p. 105111, Jan. 2022, doi: 10.1016/j.compbiomed.2021.105111.

[41] M. Pocevičiūtė, G. Eilertsen, and C. Lundström, “Survey of XAI in Digital Pathology,” *Artificial Intelligence and Machine Learning for Digital Pathology*, pp. 56–88, 2020, doi: 10.1007/978-3-030-50402-1_4.

[42] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis,” *Medical Image Analysis*, vol. 79, p. 102470, Jul. 2022, doi: 10.1016/j.media.2022.102470.

[43] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps,” *arXiv.org*, 2013, doi: 10.48550/arXiv.1312.6034.

[44] Zeiler, M.D., Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham. https://doi.org/10.1007/978-3-319-10590-1_53

[45] Springenberg, Jost Tobias, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for Simplicity: The All Convolutional Net,” *arXiv.org*, 2014, doi: 10.48550/arXiv.1412.6806.

[46] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization,” *arXiv.org*, 2015, doi: 10.48550/arXiv.1512.04150.

[47] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.

[48] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation,” *PLOS ONE*, vol. 10, no. 7, p. e0130140, Jul. 2015, doi: 10.1371/journal.pone.0130140.

[49] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *arXiv.org*, 2017, doi: 10.48550/arXiv.1705.07874.

[50] L. Shapley, “A value for *n*-person games”, **Contributions to the Theory of Games, vol.** 2, pp. 307–317, 1953.

[51] Avanti Shrikumar et al. “Not Just a Black Box: Learning Important Features Through Propagating Activation Differences”. In: arXiv preprint arXiv:1605.01713 (2016).

[52] S. Jetley, N. A. Lord, N. Lee, and Philip, “Learn To Pay Attention,” *arXiv.org*, 2018, doi: 10.48550/arXiv.1804.02391.

[53] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?,’” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16*, 2016, doi: 10.1145/2939672.2939778.

[54] Zeiler, M.D., Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham. https://doi.org/10.1007/978-3-319-10590-1_53

[55] R. C. Fong and A. Vedaldi, “Interpretable Explanations of Black Boxes by Meaningful Perturbation,” *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, doi: 10.1109/iccv.2017.371.

[56] I. J. Goodfellow *et al.*, “Generative Adversarial Networks,” *arXiv.org*, 2014, doi: 10.48550/arXiv.1406.2661.

[57] Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251. doi:10.1109/ICCV.2017.244.

[58] V. Cheplygina, de Bruijne, Marleen, and Pluim, “Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis,” *arXiv.org*, 2018, doi: 10.48550/arXiv.1804.06353.

[59] van der Maaten, L., Hinton, G., 2008. “Visualizing Data using t-SNE”. Journal of Machine Learning Research 9, 2579–2605. Available at: http://jmlr.org/papers/v9/vandermaaten08a.html.

[60] I. Gat, G. Lorberbom, I. Schwartz, and T. Hazan, “Latent Space Explanation by Intervention,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, pp. 679–687, Jun. 2022, doi: 10.1609/aaai.v36i1.19948.

[61] H. Lee, S. T. Kim, and Y. M. Ro, “Generation of Multimodal Justification Using Visual Word Constraint Model for Explainable Computer-Aided Diagnosis,” *arXiv.org*, 2019, doi: 10.48550/arXiv.1906.03922.

[62] B. Kim *et al.*, “Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV),” *arXiv.org*, 2017, doi: 10.48550/arXiv.1711.11279.

[63] M. Graziani, V. Andrearczyk, and H. Müller, “Regression Concept Vectors for Bidirectional Explanations in Histopathology,” *arXiv.org*, 2019, doi: 10.48550/arXiv.1904.04520.

[64] A. Ghorbani, J. Wexler, J. Zou, and B. Kim, “Towards Automatic Concept-based Explanations,” *arXiv.org*, 2019, doi: 10.48550/arXiv.1902.03129.

[65] Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J., 2019. This looks like that: Deep learning for interpretable image recognition, in: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, D

[66] F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning,” *arXiv.org*, 2017, doi: 10.48550/arXiv.1702.08608.

[67] tensorflow, “models/tf2_detection_zoo.md at master · tensorflow/models,” *GitHub*, May 07, 2021. https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf2_detection_zoo.md (accessed Jan. 02, 2023).