ACIT5900

# MASTER THESIS

in

Applied Computer and Information Technology (ACIT)

May 2023

Applied Artificial Intelligence

# XMask Clustering: Leveraging eXplainable AI and Clustering for Medical Knowledge Discovery

Håvard Horgen Thunold

Department of Computer Science

Faculty of Technology, Art and Design

OSLOMET

# Preface

I hereby present my master's thesis "XMask Clustering: Leveraging eXplainable AI and Clustering for Medical Knowledge Discovery." It is written as part of the Applied Computer and Information Technology master's program at Oslo Metropolitan University, where I followed the Applied Artificial Intelligence specialization. Artificial intelligence (AI) is a field in growth with what seems like endless possibilities for positive impact. That is why I decided to pursue AI studies and center my thesis around AI for medical knowledge discovery.

I would like to thank my supervisor, Hugo Lewi Hammer, for their continued support and guidance throughout this work. I am incredibly grateful for the knowledge and expertise that you have shared. I would also like to thank my co-supervisors, Anis Yazidi of Oslo Metropolitan University and Michael Riegler of SimulaMet. Lastly, I also want to express my gratitude to my fellow students and friends for engaging discussions that kept me motivated during my studies and thesis work.

Håvard Horgen Thunold
Oslo, May 2023

# Abstract

Deep Learning, a subset of machine learning, has shown great ability in supervised medical image classification tasks. Although there are significant advantages, DL models have low interpretability and are considered black-boxes. The black-box nature of these models affects trust and hinders adoption in critical domains. The field of eXplainable AI aims to address these problems by creating human-centered explanations that give insight into a model and its predictions.

This thesis answers whether the aggregation of explanations extracted from black-box models can be leveraged for medical knowledge discovery. This is done by exploring the use of explanations not only to explain the model's predictions themselves but also as a tool to reveal previously unknown properties of the data. This is done in the context of medical imaging for the purpose of extracting new medical knowledge. A novel methodology is proposed for this purpose which we call eXplanation-masked clustering (XMask Clustering). With this methodology, explanations extracted from black-box classifiers are used as masks, revealing only the areas that contributed to a prediction. This gives insight into the model's learned knowledge. Further, the masked images are clustered to uncover subclasses existing within the labeled class. Experiments with the proposed methodology resulted in explanations that accurately locate real and pseudo-real pathological identifiers. Experiments also show that XMask Clustering results in higher-quality clusters when using a combination of real and pseudo-real gastrointestinal images.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Artificial intelligence in its simplest form came to be in the late 1940s with links to theories that date back to as early as the 1800s (Schmidhuber, 2015). Since then, artificial intelligence has been applied to a wide range of applications. During the last years of the 1990s, most projects utilized what today would be called classical or traditional machine learning techniques. In the early 2000s, this changed with artificial neural networks becoming more practical as the hardware improved significantly, the amounts of data increased, and new algorithms and architectures allowed for larger and deeper neural networks. This new class of algorithms created a new subcategory of machine learning called deep learning (DL).

In DL-enabled tasks, finding patterns in large amounts of unstructured data is important. One such task is the classification of medical images. DL has been shown to accurately detect identifiers of disease in various medical imaging such as chest radiology imaging, mammograms, and magnetic resonance imaging (Senaras & Gurcan, 2018; Tahmassebi et al., 2018). To accurately make such predictions, the DL model extracts medical knowledge from the images and associates this knowledge with the pre-defined labels. This is limiting in that the model only makes classifications based on the labels given during the learning phase (existing medical knowledge). One can, however, build upon a DL model's ability to extract medical knowledge from images by applying methods of knowledge discovery on top of the knowledge extraction components. This has the potential to bring new medical discoveries while being time and resource-saving by moving away from traditional methods that rely on hand-crafted features (Zhang et al., 2022). This is highly relevant for data coming from medical institutions that are not manually labeled. The data may be retrospective data, with some data simply marked as disease and others not. This could be from routine checkups such as a colonoscopy exam. In such cases, it is unknown which characteristics of the images resulted in the separation of healthy and diseased. Automatically finding these characteristics and leveraging this for knowledge discovery using DL is a challenging and less researched area and is, therefore, the focus of this thesis.

Although there are potential benefits, there are also downsides that come with using DL models for knowledge discovery that will be addressed in this thesis. DL models are inherently difficult to interpret and are often referred to as black-box models. This can pose a problem in industries where AI is used to make critical decisions. In the medical domain, there must be appropriate levels of trust in the AI system, both from the medical experts and the patients. To solve these problems of black-box models, a new field called eXplainable Artificial Intelligence (XAI) has been created (Adadi & Berrada, 2018). XAI provides techniques for creating explanations of DL models and their predictions. This can be visual explanations showing which part of the image had the most influence on the AI's prediction. In the context of medical imaging, such explanations show the model's learned medical knowledge. Ribeiro et al., 2016 showed that visual explanations also adjust a user's trust in the model to a level that is more appropriate relative to its performance.

In this thesis, we provide a comprehensive look into how these explanations can be utilized by unsupervised training algorithms to provide a more complete insight into the medical data in which the DL method was trained on. In other words, this thesis explores the use of explanations not only to explain the model predictions themselves but also as a tool to potentially reveal previously unknown properties of the data. This is done in the context of medical imaging and with the aim of extracting new medical knowledge from such data. A new methodology is proposed for this purpose which we call eXplanation-masked clustering (XMask Clustering).

## 1.1 Goals & Objectives

The goal of this thesis is to answer the following two research questions. First, can explanations of a black-box model be leveraged to reveal new medical knowledge while addressing the problems hindering the adoption of deep learning in the medical domain? Second, is using pseudo-real medical data, adding a synthetic layer on top of real medical images, useful for the evaluation of such a technique? To answer these questions, the following objectives have been identified:

**Objective 1** Research the current state of deep learning for new medical knowledge discovery and any links to XAI in the existing literature.

**Objective 2** Develop and implement a methodology for new medical knowledge discovery using visual explanations extracted by XAI techniques.

**Objective 3** Evaluate the performance of the method using a combination of pseudo-real and real medical data.

## 1.2 Main Contributions

The research conducted in this thesis aimed to provide valuable insights and contribute to a highly unexplored area of research. By experimenting with new ideas for XAI, medical knowledge discovery, and data, we have received insights that we believe are of importance to the field. The contributions can be summarized for each research question (RQ) from Section 1.1.

**RQ 1** *Can explanations of a black-box model be leveraged to reveal new medical knowledge while addressing the problems hindering the adoption of deep learning in the medical domain?*

A novel methodology is proposed called XMask Clustering. The methodology is shown to accurately uncover subclasses of medical image data by leveraging the information revealed by XAI for improved clustering. Baked in XAI forces interpretability, increased trust, and model insights and thereby addresses some of the problems that hindered the adoption of black-box models in the medical domain.

**RQ 2** *Is using pseudo-real medical data, adding a synthetic layer on top of real medical images, useful for the evaluation of such a technique?*

A pseudo-real medical dataset is created using the gastrointestinal dataset HyperKvasir at its base. Experiments showed that pseudo-real data is suitable for evaluating visual explanations and revealing errors that may be difficult to find when using real medical data.

## 1.3 Thesis Structure

This thesis is structured using seven chapters. Immediately following the introduction is Chapter 2, which provides details on the concepts necessary for understanding the technological base that this thesis builds upon. Chapter 3 presents a review of related research, covering deep learning for knowledge discovery and eXplainable AI in the medical domain. Chapter 4 details the data and experiments. Chapter 5 then presents the results of the experiments along with an interpretation and evaluation. Chapter 6 gives an analysis of the work along with a discussion on ethical considerations. Lastly, future work and some concluding remarks.

# Chapter 2

# Underlying Concepts

This chapter covers important concepts used in this work. In order to get a complete understanding of the literature and research conducted in this thesis, it is essential to have a thorough understanding of these underlying concepts. Advanced readers could consider jumping to the related work in Chapter 3.

## 2.1 Machine Learning

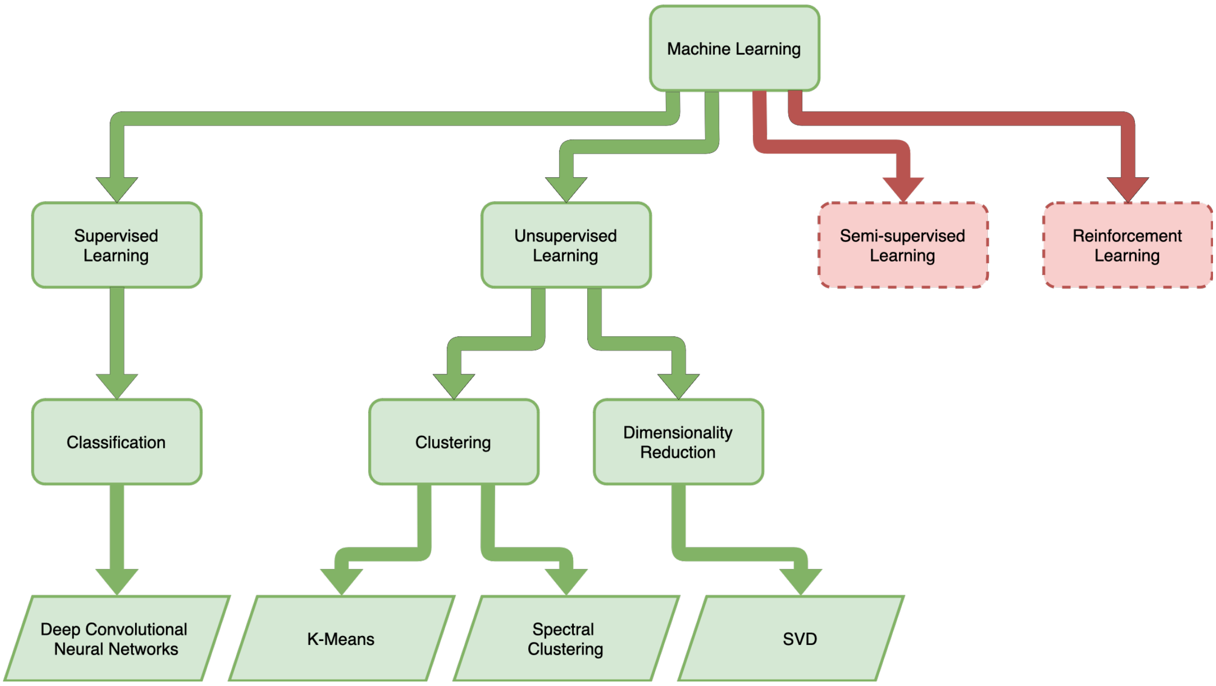

Figure 2.1: High-level overview of a small subset of Machine Learning concepts with green indicating concepts relevant to this thesis and red-dashed rectangles indicating non-relevant concepts.

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) where the goal is to build computer programs that can learn from experience, in the form of data, and make predictions based on what has been learned (Mitchell, 1997). Machine Learning can be categorized into four main categories: (1) supervised learning, (2) unsupervised learning, (3) semi-supervised learning, and (4) reinforcement learning (Das & Behera, 2017). In supervised learning, the data is labeled meaning that each data point includes information about the class to which it belongs prior to learning. A model is then trained to best fit the data points to the labels. Unsupervised learning, on the other hand, does not use labeled data. In unsupervised learning, machine learning algorithms are used to learn patterns and groupings in the data without requiring human-made labels. Semi-supervised learning combines the two approaches by using both labeled and unlabeled data. The last major category is reinforcement learning where the data comes from interactions with the environment and learning occurs based on a reward signal. This thesis focuses on supervised and unsupervised learning with further sub-categories as seen in Figure 2.1.

Going deeper into the hierarchy of machine learning and looking at the evolution of ML, hardware, and data, a broader category of ML was born called deep learning (DL). DL can be defined by the number of layers in an Artificial Neural Network (ANN). For an ANN to be considered deep it must have three or more layers although other definitions also exist (Schmidhuber, 2015). Deep Neural Networks (DNN) combined with more data has shown to perform better than classical ML algorithms (Dong et al., 2021). DL also differs from classical ML in that feature extraction is done by the model itself. It has revolutionized the field and has enabled applications in speech processing, natural language processing, computer vision, generative AI, and more. DNN models do however come with the caveat that they are black-boxes meaning they are inherently uninterpretable. This spawned a new field called explainable AI (XAI) with the goal of explaining these black-boxes and their predictions. As this thesis focuses on using such black-box Deep Neural Networks (DNN) trained in a supervised manner for image classification in the medical domain, this section starts by providing background on supervised image classification, architectures important for this work, and training and evaluation concepts. This is followed by concepts related to unsupervised image classification to provide the necessary background for understanding how ML can be used to identify new sub-labels in data. Further, a key part of this work is extracting human-centered explanations from the black-box image classifier. This section, therefore, completes by detailing the concepts of XAI that are most important to this thesis.

## 2.2 Supervised Image Classification

Image classification is the task of learning to assign a label to an image. In this section, there is an assumption of the existence of labels. This thesis also explores how unknown sub-labels can be identified. Finding these different groupings of images in an unsupervised manner is called clustering and is detailed in Section 2.5. Assuming a set of image and label pairs a model can be trained in a supervised manner to best predict the labels of images that are new to the model. This has use cases in a wide range of domains, such as the medical domain where image classification can differentiate between healthy and pathological images. The implementation of such an image classifier could be done using a fully-connected DNN, but this would require an extreme amount of parameters when handling images. As image sizes increase it becomes infeasible to use a fully-connected DNN for this task. To efficiently learn to classify images, a more efficient architecture such as the Convolutional Neural Network is needed.

### 2.2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNN) for classification traditionally consist of convolutional layers, pooling layers, and fully connected final layers called classification heads. The convolutional layers are the heart of a CNN and take advantage of the mathematical convolution operator. An $nxn$ pixel window called a kernel slides across the input extracting features that are local to the kernel window resulting in feature maps. The weights of the kernel are learned with training of the CNN. This is followed by the pooling layer. This is a layer that serves two main purposes. The first is that pooling reduces the model's sensitivity to a feature's position. This is called local translation invariance. The other is downsampling of the feature map by running a pooling operation such as taking the maximum or the average value. A similar effect to that of max pooling has been shown to occur if one uses a convolutional layer where the kernels are moved across the image with a stride greater than or equal to 2 (Springenberg et al., 2015). By applying necessary padding and using a stride of 2, the image will halve in size, and over multiple layers, the input will be reduced significantly. The result is then fed into a fully-connected layer where the last layer has $K$ neurons where $K$ is the number of classes in the dataset. This gives a prediction for each $K$ classes and forms a basic image classifier. For state-of-the-art image classification, the Convolutional Neural Network and its convolutional layers have become a standard backbone of many more advanced architectures such as VGGNets, ResNet, Inception, and Xception. For this thesis, the ResNet architecture is used and is therefore further detailed in the next section.

### 2.2.2 ResNet

Deep convolutional neural networks have a problem: there is a point where increasing the number of layers causes a degradation of training accuracy and an increase in training error. He et al., 2016 introduced Residual Networks (ResNet), an architecture for deep networks that allow for an increased amount of layers without the degradation problem. ResNet is a CNN-based architecture that introduced the concept of a residual block and skip connections. A residual block is a stack of two or more convolutional layers with a skip connection that connects the input of the first layer in the block with the output of the last layer. A ResNet is a stack of these residual blocks with average pooling and a fully connected layer forming the final two layers. The authors showed that ResNets have improved accuracy as the depth increases and do not suffer from the same degradation problem that plain networks do when tested on two major datasets.

## 2.3 Supervised Image Classification - Training

The training of an image classification model in a supervised setting is an optimization problem where the aim is to find the minimum loss or error between the true values and the predicted values. Commonly gradient descent or a variant of gradient descent is used as the optimizer. This is an iterative method that updates the values of the network in the opposite direction of the gradient. This is the derivative of the loss with respect to the weights in the network. By applying the chain-rule one can get the derivatives with respect to weights all the way back through the network allowing the network as a whole to be updated based on the loss, this is called backpropagation (LeCun et al., 2012). This is the basis of training ML models, but there is a multitude of other concepts that play a role and need to be considered, such as the initialization of the network, the selected optimizer, activation functions, generalization, and regularization.

### 2.3.1 Optimizers

Since the inception of gradient descent, many variants have been created such as Adagrad, RMS Prop, and Adam. These optimizers are smarter in the sense that they provide functionality for variable rates of learning. This gives significant improvements to training speed by allowing for larger changes to occur when far away from the target and smaller changes when close to the target. Although there is no one optimizer that is best in all scenarios, Adam is currently recommended as a default optimizer for many deep learning problems (Godbole et al., 2023).

### 2.3.2 Activation Functions

Activation functions are mathematical functions applied at the output step of each neuron in an ANN with the purpose of bringing nonlinearity to the network. Nonlinearity is required in order to solve real-world problems which are often nonlinear (Dong et al., 2021). A multitude of activation functions have been proposed such as logistic sigmoid, tanh, and variants of rectified linear units (ReLU) (Dubey et al., 2022). Sigmoid and tanh were traditionally used extensively, but the saturated gradients at the two tails showed to be problematic in training deep neural networks. Variations of these activation functions, such as scaled hyperbolic tangent, scaled sigmoid, and parametric sigmoid, have therefore been made as an attempt to solve these problems. Although these variants minimize the problem, the vanishing gradient problem can still be observed. Rectified activation functions, however, do not have the same problems. ReLU is a simple and computationally efficient activation function that is linear for positive values and zero otherwise. Negative values always being zero may cause a problem where neurons become inactive and only output zero, this is called the dying ReLU problem. A solution is to use leaky ReLU where a slope is added on the negative side which allows for utilization of the negative values (Maas et al., 2013).

### 2.3.3 Initialization

An ANN needs an initial set of values for the weights in the network. How these values are initialized can affect how quickly a model will converge or if it will converge at all. Initializing too low will result in minimal changes to the network due to the vanishing gradient problem. Initializing too high will results in the inverse called the exploding gradient problem. Uniformly and normally distributed random has also been shown to not be ideal. The solution is a more specialized initialization method. When using the tanh activation function Xavier initialization, also called Glorot initialization, has shown good results (Glorot & Bengio, 2010). It ensures that the mean of the activations are zero and the variance is constant throughout the layers. This prevents vanishing and exploding gradients. Another initialization method that has shown similarly good properties when using ReLU activation function is He initialization (He et al., 2015).

### 2.3.4 Generalization and Regularization

During the training process of a deep neural network, one may encounter overfitting. Overfitting refers to when the model too closely focuses on the training data and thereby does not perform well on new data. In other words, the model does not generalize well. Methods for reducing overfitting are called regularization techniques (Goodfellow et al., 2016). There are a host of methods in the literature that try to combat this problem such as increasing the amount of data through data augmentation, dropout, batch normalization, and early stopping.

A model will more easily learn the noise and random fluctuations in the data, rather than the general pattern, in small amounts of data (Li et al., 2019). The amount of training data can therefore be seen as crucial in preventing overfitting. One technique for increasing the amount of data available under training is data augmentation (Shorten & Khoshgoftaar, 2019). In computer vision applications, data augmentation can be simple augmentations such as flipping, rotating, shifting, and scaling images. It can also include more advanced techniques such as using Generative Adversarial Networks to create new data (Perez & Wang, 2017).

For neural networks, dropout is a simple and effective method for reducing overfitting (Srivastava et al., 2014). Dropout effectively changes the model's architecture by randomly selecting neurons that will not be part of an iteration. This prevents neurons from learning the noise in the training data and improves generalization. For Convolutional Neural Networks there is evidence that dropout can be replaced by using batch normalization layers (Garbin et al., 2020). Batch normalization normalizes the inputs of all layers which has been shown to be effective in improving training efficiency by allowing for larger learning rates to be used. It has also been shown to have a regularization effect.

Another simple regularization technique is early stopping. Continued training increases the model's accuracy on the training set, but there comes a point where there is no improvement on the validation or test set. Further training results in overfitting. A simple countermeasure is therefore to stop training when there has been no improvement for a select number of iterations.

### 2.3.5 Transfer Learning

Modern neural networks require a large amount of data to be trained. In some domains, such as the medical domain, getting large amounts of data may not be possible (Waisberg et al., 2023). Transfer learning can be a solution to this problem across many domains with this work focusing on transfer learning for computer vision. Transfer learning is a research area where the premise is to use the knowledge that has been learned on one task and transfer that knowledge onto another task where you may have a limited amount of data (Brodzicki et al., 2020). In transfer learning, one selects a source model that is pre-trained on a source dataset. This model is then used as the source for training the target model in the target domain using the target dataset. Further one must decide on one of two strategies for transfer learning: (1) feature extraction, and (2) fine-tuning. With the feature extraction strategy all layers are frozen, meaning that they are not updated during backpropagation, while the classification head is replaced and trained to fit the target task. This strategy is appropriate when the target dataset is small. Another strategy is that of fine-tuning. With fine-tuning one or more layers are unfrozen. Here the pre-trained weights are serving a similar purpose to other network initializations but give a significantly improved starting point. This is most often an appropriate method if the target dataset is of sufficient size to train the entire network. Larger data sizes are needed to train larger networks and with transfer learning as the size of the data grows the returns on using transfer learning diminish logarithmically (Sun et al., 2017). Further research on transfer learning for computer vision with large, small, and highly detailed datasets, has revealed more elaborate methods.

Cui et al., 2018 proposed a transfer learning method where the pre-trained network is trained on a specifically calibrated source dataset. This stands in contrast to the more traditional method of using a pre-trained model trained on a large and general source dataset. In their work, they use ImageNet (Russakovsky et al., 2015) as the basis of their source dataset containing millions of images across thousands of classes. For their target dataset, they used the iNaturalist (iNat) dataset containing fine-grained images of nature (Van Horn et al., 2018). To further select the subset for use in the training of the source model, they calculate the Earth Mover's Distance between the ImageNet images and their iNet target dataset. Two subsets were used in their experiments taking the top 20% and 40% most similar classes from ImageNet. They show that using these subsets can increase the accuracy of the target model by about one percentage point. Another key takeaway is that transfer learning from a less detailed source is beneficial also when targeting fine-grained domains. Although it gave an overall performance increase, some classes of the target dataset were negatively impacted.

Research focusing on transfer learning in medical imagining has shown similar results. Alzubaidi et al., 2020 experimented with using medical images of one domain as the source and medical images of another domain as the target. An experiment outside of the medical domain using images of animals was also made for comparison. Their experiments showed performance increases when the source and target datasets are in the same domain. Their experiments also showed that a very large source dataset, such as ImageNet, that is slightly related to the target dataset is better than using a small source dataset that is closely related to the target dataset. A large general dataset can therefore be a good alternative if one is not available in the target domain. Another paper using transfer learning in the medical domain showed that using a source dataset that visually seems similar to the target dataset may not provide the best results indicating that the importance of having visually similar source and target datasets is task dependent (Moran et al., 2021). Both source datasets tested do, however, show improved results over not using transfer learning at all. A common conclusion on the importance of similarity between the source and target dataset is not directly present. One thing that is shared across all of the research, however, is that using transfer learning is beneficial to the performance of the model although with fewer benefits being seen as the target dataset increases in size.

## 2.4 Supervised Image Classification - Evaluation

Evaluating a Machine Learning (ML) model is the act of measuring the performance of a model on a given task using estimation techniques such as K-fold cross-validation and quality metrics like accuracy, precision and recall, or F1-score (Stąpor, 2018). ML models should be evaluated continuously throughout training using an appropriate metric and dataset. Generally, when training an ML model the dataset is split into a training set, a validation set, and a test set. The model is evaluated using the validation set and is used to further tune any hyperparameters. The validation set can leak into the model and is therefore not independent and cannot be used to get an unbiased evaluation of the model's performance. The test set sometimes called the hold-out set, is therefore used for this purpose. An improvement over this is to use K-fold cross-validation. This is a resampling technique where buckets of training and validation data are created K-times. The model is then evaluated on all K-buckets of validation data. This provides a more representative picture of the data and a less biased evaluation of the model's performance. How the performance is measured is using a metric. For classification problems, which are the focus here, performance is often measured using accuracy, precision and recall, or F1. Accuracy is the most simple representing the amount of correctly classified samples over the total samples. This poses a problem in cases where the data is not balanced as skewed data will also skew the accuracy. An accuracy of 95% does not represent a high-performing model on a dataset where one class represents 95% of the data. A better metric in such cases would be to use precision and recall or F1-score. Precision represents true positives over predicted positives and recall represents true positives over actual positives. This means that for cases where false positives must be avoided at all costs, one should focus on high precision and low recall while in cases where false negatives should be avoided one needs to focus on high recall and low precision. The F1-score can simplify this by creating a single number from precision and recall.

## 2.5 Unsupervised Learning - Clustering

Clustering is an unsupervised machine learning technique where the task is to find groupings in unlabeled data. It has a wide range of use cases, such as anomaly detection, compression, or revealing interesting properties of image data. In this work, the focus is on clustering for image classification using Spectral clustering and K-Means.

### 2.5.1 Spectral Clustering

Spectral clustering is a technique for improved clustering that comes as a consequence of The Curse of Dimensionality (Ng et al., 2001). That is, the problems arising when working with high-dimensional data, are apparent when working with image data. Clustering a flattened version of the image directly is not practical. Common practice is therefore to cluster on features extracted from the images. This can be done using traditional manual methods for shape retrieval, such as extracting edges using a Prewitt kernel, or color features using a color histogram. A more modern approach is to leverage the automatic feature extraction of deep convolutional neural networks. The output of a convolutional layer can be used to get a lower-dimensional feature vector representation of an image. This feature vector may still be larger than one wants. With spectral clustering, dimensionality reduction techniques are used prior to applying the clustering algorithm. Spectral clustering is thereby not a clustering algorithm in itself, but rather an extra step for improved efficiency and reduced computation time. An algorithm such as K-Means can then be used to cluster this lower-dimensional data.

### 2.5.2 K-Means and X-Means

K-Means is a simple clustering algorithm with a time complexity of O(n) using big-O notation. The algorithm starts by initializing a centroid for each $K$ clusters. There are various strategies for this initialization. A simple strategy is to take $K$ random points from the dataset and use these as the initial centroids. This step causes K-Means to be dependent on a good initialization for convergence. To minimize this issue one can run K-Means multiple times. Another strategy called K-Means++ has shown to be a more robust initialization scheme that outperforms standard K-Means in both accuracy and time (Arthur & Vassilvitskii, 2007). K-Means++ chooses the initial centroids by using a probability that is based on a point's distance to the current centroids. With the initialization in place, all points are assigned to their nearest centroid. The algorithm's main loop is then run by shifting the centroids towards the mean of the points attached to the centroid and re-assigning each point to the nearest centroid. K-Means can also be formulated as an optimization problem that minimizes the squared distance between points and centroids for each cluster.

One downside of K-Means is that $K$, the number of clusters, must be defined. When the number of clusters is unknown, one can use a technique called X-Means. With X-Means, the K-Means algorithms are run with different values for $K$. The number of clusters that are most appropriate for a dataset can then be found by analyzing various clustering performance metrics.

## 2.6 Unsupervised Learning - Clustering Performance Evaluation

Clustering performance evaluation metrics can be split into two main categories, those requiring labeled data and those that do not. For this work the most important of which are Rand Index, Silhouette Coefficient, and Davies-Bouldin Index (Fahad et al., 2014; Shutaywi & Kachouie, 2021).

Rand Index is a way of measuring the similarity between the labels that the cluster has assigned to data points and the ground truth labels. This differs from standard accuracy measures as with clustering the label of the cluster that a data point is assigned to may not be the same as its label. To properly measure the accuracy of a clustering one must therefore ignore permutations. Rand Index can properly give a score in the range $[0, 1]$ indicating the number of matching pairs between the cluster labels and ground truth. It is highly interpretable, but when there are no labels available other methods must be used.

Silhouette Coefficient is a metric that can be used when no labels are available. It gives a value that increases as the clusters become more defined. More defined here refers to the distance between the points within the cluster being low while the distance to other clusters is high. It gives a value that is easily interpretable in the range $[-1, 1]$ where $-1$ indicates an incorrect clustering while $1$ indicates the highly dense clusters that are well separated.

The Davies-Bouldin Index differs from Silhouette Coefficient in that it focuses less on the quality of the clusters and more on the separation of clusters. A low Davies-Bouldin index indicates a high degree of separation between clusters with 0 being the lowest possible value with no upper bound.

The within-cluster-sum-of-squares, known as inertia, is a calculation of the distance between points in a cluster and its mean point. Lower values indicate better clusters, but it does give large values to elongated clusters even though they may be of high density.

## 2.7 Explainable AI

Explainable AI (XAI) is a field of artificial intelligence that aims to give deeper insights into black-box models and their predictions. Researchers have cited trust, performance, legal (regulation), and ethical considerations as reasons for XAI (Barredo Arrieta et al., 2020). This has become increasingly important as the adoption of AI has reached critical areas such as the medical domain. How this is done for models that are not inherently interpretable through external XAI techniques.

External XAI techniques can be explanations of single predictions in the form of text or visualizations, or explanations of models in their entirety using examples, local changes, or transformations to more interpretable models. Text and visualization explanations provide a direct and human-understandable explanation, but normally only for a specific prediction. Using examples can provide a more general understanding of a model by providing similar examples and predictions to the prediction that one wants to understand. This does, however, not provide a direct explanation for a prediction of interest. Local explanations look at a subset of the problem and attempt to provide explanations within that simplified context. The last method is leveraging more interpretable models either by using a mimic model where the mimic model is an interpretable model that learns the behavior of the black-box model or by replacing the black-box model entirely. In this thesis, visual explanations are used.

Hoffman et al., 2018 provide definitions for four techniques that can explain image classifiers by altering the input. The first technique is the method of concomitant variation where the authors suggest the use of image processing filters to create contrast cases. The altered image is then used as input to the model and any variation in the model output can then be assessed by a human observer. The method of agreement is the second technique. This technique explains the model's behavior by checking the degree of agreement between the model and the human. The human must with this technique hand-craft images where parts of the image, that the human believes should be of importance in the classification, are removed. When this image is given to the model a drop in confidence is expected symbolizing that the model and the human agree on the importance of the removed part for the given class. The third technique is the method of difference. Here a correctly classified image is fused with an image that the model should not know to classify. A decrease in prediction confidence is expected and can reveal a model's inner focus by observing the parts of the original image that was changed. With an increase or flat confidence, one must ask which parts of the overlaid image gave the model an indication of the original image class. This technique could also lead to fused images where the original class is no longer identifiable by a human. In such cases, the model should, similarly to a human, not be able to identify the image. The fourth and last technique proposed is the method of adjustment. The goal of this technique is to evaluate a model's robustness in seeing an unknown image. This can provide a further understanding of the model such as the degree to which the context surrounding an object is used.

The technical implementation for extracting visual explanations from an image classifier contains two parts: (1) an attribution algorithm that provides the data for the explanation and (2) the visualization which uses the data to show a human-understandable explanation. Attribution algorithms for image classification can generally be split into two categories: (1) gradient-based methods and (2) occlusion-based methods. Gradient-based methods look at a specific layer in the model and calculate the influence of an input feature on the output of the layer by applying tiny changes to the input. This method allows for the evaluation of any given layer in a model and can provide a deeper understanding of the inner workings of the model, such as the convolutional layer that has the highest influence on the prediction. Occlusion-based methods apply squares, often gray or black in color, onto the input image and record any changes in the prediction. The process is repeated until the entire image has been occluded. The result is an occlusion sensitivity map that shows the importance of each pixel. This is a human-understandable explanation that does not expect the same expertise of the end-user as with the gradient-based methods.

# Chapter 3

# Related Work

This chapter presents relevant research with the aim of building a context for which this thesis fits. First focusing on a core component of this work, eXplainable AI (XAI), and its use in the medical domain. Lastly, the role of AI and XAI in medical knowledge discovery.

## 3.1 Explainable AI

Explainable AI (XAI) techniques targeting image classifiers contain two parts: (1) the attribution algorithm, and (2) visualization. This section looks at what the current knowledge is for each part for general use cases. This is further narrowed down to XAI in the medical domain in the next section.

### 3.1.1 Occlusion-Based Methods

Occlusion-based methods, also called perturbation-based methods, were first introduced by Zeiler and Fergus, 2014. In their paper, they showed that by systematically applying a gray square to the input image of a CNN they could create explanations for individual classifications. The explanation is an occlusion-sensitivity map which contains the confidence difference between the original image and the occluded image for each pixel. Further, they use the occlusion-sensitivity maps to show that a properly trained CNN focuses on the object and only uses the surrounding context to a minimal extent. They did so with state-of-the-art networks and multiple datasets. Although the authors exclusively used CNNs in their research, the proposed method of explanation is model agnostic as it only relies on altering model input and recording model output. The proposed method does, however, also come with one main downside, computation cost. The authors do not cover this limitation, but later research has identified it as a problem.

Randomized Input Sampling for Explanation of Black-box Models (RISE) proposes an occlusion-based method to tackle the large computation previously associated with occlusion-based methods (Petsiuk et al., 2018). With the original method, a square mask is moved across the image with some stride defining the amount of movement. With a small mash and a small stride, this will result in a large number of masked images and thus a large number of inferences required for a single explanation. RISE attempts to solve this by applying $N$ random masks where $N$ is a hyperparameter. The masks are upsampled using bilinear upsampling providing smoother edges in the mask and thus smoother occlusion sensitivity maps. The resulting explanation is a weighted sum of the difference in prediction by the black-box model, here a ResNet50 model. The authors showed that this requires significantly less computation than the previous method by Zeiler and Fergus, 2014, but still higher than gradient-based methods. RISE does, however, have a limitation that is inherent in its random nature, it cannot guarantee that all parts of the image will have been masked as part of the $N$ masks.

Building on the work of RISE is Morphological Fragmental Perturbation Pyramid (MFPP) (Yang et al., 2021). Similarly to RISE, they apply occlusion randomly, but they do so to segments (superpixels). These segments are used to leverage the full morphology of the objects in the image. The authors argue that previous work only takes advantage of two of three identifiers of objects, color and texture, while morphological theory says that shape is just as important. This makes the patches of previous work less ideal. Superpixels, on the other hand, can align with the edges of objects. MFPP occludes these segments and creates occlusion sensitivity maps in the same manner as previous work. With this base, the authors extend their method by using a pyramid of segmented images where each image uses a different scale of segmented regions. This can catch features at different levels of detail. The authors argue that the use of superpixels allows their method to take full advantage of the morphology of the classified objects and combined with the hierarchical segmentation can provide explanations in a computationally efficient manner.

Another attempt at solving the problem of computational cost for occlusion-based methods is Hierarchical Occlusion (HihO) (Monroe et al., 2021). Their method relies on capturing features that cover a large space rather than creating explanations that exaggerate small features. They apply occlusion by setting zeroing pixels covering the full length of the image in one dimension and half the length in the other dimension. The size of the occlusion is always the full length in one dimension while it halves for each iteration in the other dimension. The method is executed until the change in the resulting confidence difference is lower than a predefined threshold. The authors have found that this method requires orders of magnitude fewer iterations than other occlusion-based methods. They cite a difference of 65 thousand iterations versus their 12 iterations. It should be noted, however, that they use a naive approach when representing the original occlusion-based method by using a stride of 1. This would have given an explanation that is significantly more detailed than that of the proposed HihO method.

Fong and Vedaldi, 2017 takes a different approach than the previously mention occlusion-based methods by applying blurring to represent a deleted part of the image. Their approach is to learn to find the minimal region of the image that needs to be blurred in order to reduce the model's confidence by 99%. By focusing on minimal occlusion, the authors argue that their explanations can visualize what differentiates the classified object from other objects in the dataset. They showed that this can give highly detailed occlusion sensitivity maps on common datasets in computer vision.

### 3.1.2 Gradient-Based Methods

Gradient-based methods are another set of attribution methods that, as the name implies, utilize gradients to measure the importance of changing the model's input. The first such method, within the field of image classification, was proposed in the paper "Deep Inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps (Simonyan et al., 2014)." Similarly to the first paper on occlusion-sensitivity maps, the goal was to explain a CNN. The difference, however, is that the gradient-based methods are not model agnostic. The proposed method is to take the derivative of the output (model's confidence for the class in question) with respect to the input. In short, the method is similar to backpropagation in training but takes the derivative with respect to the input rather than the weights of the network. The result is a gray-scale image where individual pixels are lit-up indicating their importance to the classification. These images do, however, include quite a bit of noise.

Sundararajan et al., 2017 proposed integrated gradients (IG) as a new attribution method for improved visualization and better insight into a model. The first step of their method is to create a linear interpolation between a baseline image, such as a zero-valued image, and the target image. The number of steps in the interpolation is defined by a parameter $m$. Gradients are then calculated $m$ times and summed. The authors empirically show that this gives better explanations, compared to earlier input-gradient methods, across multiple network types. This does, however, come at the cost of computation time with the authors estimating between 20 to 300 calls. This is still significantly lower than that of some occlusion-based methods.

Further work on integrated gradients aims to reduce noise found in explanations. Guided Integrated Gradient (GIG) by Kapishnikov et al., 2021 proposes adjusting the straight-line path from IG. Similarly to IG, there is an interpolation from the baseline to the input explanation. The difference is that with IG every pixel is moved at every step while with GIG only the 10% of pixels that have the lowest partial derivative are moved to be closer in value to that of the input. A step parameter $m$, like with IG, is used to define the amount of movement. A larger number of steps, and thus higher computation needs, results in more accurate explanations. Visually GIG results in explanations that have less noise than IG.

Another attempt at reducing noise in explanations from gradient-based explanations is SmoothGrad (Smilkov et al., 2017). The authors argue that the direct use of gradients as explanations of the importance of features (pixels in the case of images) can be misleading as a strong feature may saturate the gradient. They also argue that networks commonly use ReLU activation functions that will have fluctuating derivatives causing noisy explanations. The proposal is therefore to take a local average which will result in a smoother result. A Gaussian random sample of the neighborhood is taken as direct computation would be costly. Using a CNN-based network and common image datasets, they show that this method produces less noisy explanations -compared to vanilla input-gradient and integrated gradient- that more directly match that which is expected by a human observer.

Other work on gradient-based methods takes a different approach to visualization that is close to that of the occlusion-based methods. Grad-CAM proposes taking the gradients at the last convolutional layer of the network, rather than the output of the network seen in previous work, and using this as the basis of the explanation (Selvaraju et al., 2017). They argue that the spatial information preserved by the convolutional layer is not found in the fully-connected classification head and therefore makes it a better choice. Further, Grad-CAM requires global average pooling as the penultimate layer. The resulting Grad-CAM explanations highlight the regions that have the most effect on the prediction. The idea of creating visualizations by taking gradients at the convolutional layer is built upon by FullGrad (Srinivas & Fleuret, 2019). FullGrad aggregates the gradients across all convolutional layers to create the final visualization. They show that this provides sharper visualizations that are also more accurate.

### 3.1.3 Visualization

Visualizations are the resulting explanations of the attribution methods detailed in the previous sections. In the literature, there are conflicting opinions on what constitutes a good explanation. The authors of Grad-CAM define a good explanation as an explanation that provides justification for a classification by localizing the parts of the image that contributed the most to the given classification (Selvaraju et al., 2017).

Further, they say that the explanation must be of an appropriate resolution. These conditions for good explanations are vague and open to different interpretations. With the occlusion-based methods suggestions for decreasing computation time came at the cost of a more rough visualization showing larger regions rather than smaller attribution details. Here the authors focused more on the first point of a good explanation. The gradient-based methods showed problems with noise in the resulting visualizations. These explanations are of high resolution, but may not provide the expected justification for the classification as the authors of SmoothGrad see only a rough overlap between these explanations and a human's expectation (Smilkov et al., 2017). To create visualizations more aligned with expectations, SmoothGrad proposes smoothing values by taking a local average and capping values to produce more intelligible visualizations at the cost of correctness. Srinivas and Fleuret, 2019 suggests that the trade-offs seen in visual explanations are an inherent limitation of 2D visualizations. They argue that limitations and trade-offs made must be explicitly stated when using visual explanations as a 2D visualization will never fully represent a model.

Other works on improved visualization suggest that previous work has focused too heavily on what parts of an input image change the classification for the positive without considering how the change affects in the negative direction (Rudin, 2019; Zintgraf et al., 2017). The argument is that both are needed to provide a user with a full explanation and are necessary for the adoption of AI in critical fields.

## 3.2 Explainable AI in the Medical Domain

A significant amount of research on XAI points to the advancements of deep learning and its applications in the medical domain as reasons for XAI (Kapishnikov et al., 2021; Petsiuk et al., 2018; Smilkov et al., 2017; Sundararajan et al., 2017; Yang et al., 2021). The research introducing these XAI methods does, however, not use datasets from the medical domain but rather benchmark datasets for computer vision. This section, therefore, reviews research on how XAI has been used in the medical domain.

Gecer et al., 2018 used XAI as a sanity check, to ensure that their black-box model's understanding matches their expected understanding. They used an occlusion-based method similar to that of Petsiuk et al., 2018 with randomly placed occlusion patches. This was run on their cancer classifier for breast histopathology images. They did, however, not have the necessary expertise to say whether the visualization provided is correct from a medical perspective. Papanastasopoulos et al., 2020 also used XAI for a similar purpose showing that gradient-based methods can be useful in understanding the lower-level features learned by a model when trained on medical images. This is a common usage of XAI in research on the classification of medical images (van der Velden et al., 2022).

Xie et al., 2020 takes this one step further and uses the explanations to create alternate visualizations meant for a physician. They train a binary classifier on normal and abnormal brain ultrasound images and use Grad-CAM, described in Section 3.1.1, to provide explanations for abnormal images. The explanations are then used to draw bounding boxes on top of the original image to localize the abnormalities. The authors deem this useful but also identify that Grad-CAM's accuracy is a limitation as parts of the abnormalities lie outside of the highlighted areas. This method has similar aspects to the first few steps of the work presented in this thesis in that it uses a binary classifier and then focuses on localizing abnormalities using explanations. It does, however, stop prior to the knowledge discovery step which is central to this thesis.

Other works on XAI in the medical domain look at how XAI can reveal hidden bias. Mahmoudi et al., 2022 used both gradient and occlusion-based methods to evaluate a model trained to detect Covid-19. They used multiple state-of-the-art architectures and fine-tuned from ImageNet to the target X-ray images. They showed that XAI can reveal bias as all XAI visualizations showed a strong focus on parts of the image that should not be used for the classification. This is with a model achieving 96% accuracy and showed that one cannot select a model purely based on performance metrics as the model may have hidden biases. Given that the explanations showed a focus outside of the lungs, which is the area of focus, the researchers segmented the lungs prior to further analysis. This limited the biases.

Although the reviewed work has shown that there are benefits to using XAI in the medical domain, work also shows that it is important to select the appropriate attribution method and visualization for the task. Guided Integrated Gradients showed an improvement in visualization as compared to previous gradient-based methods on ImageNet, but there is a downside noted by the authors. The downside comes with the removal of noise that is central to the method. For medical imaging, where identifiers of disease may be small and spread throughout the image, GIG may remove important identifiers (Kapishnikov et al., 2021).

The importance of attribution method selection is also evident in other research. Ehrhardt et al., 2019 found that gradient-based methods, such as guided integrated gradient and Grad-CAM, provided poor explanations on pathological retinal OCT and brain lesion MRI images while occlusion-based methods resulted in more plausible explanations. The Guided Integrated Gradient and Grad-CAM papers both argued that their XAI method provides improved explanations, but this was on the ImageNet dataset. Ehrhardt et al., 2019 showed that this result was not transferable to the medical domain.

## 3.3 New Medical Knowledge Discovery

This section looks at the role of artificial intelligence in new medical knowledge discovery. Literature on AI-based drug discovery is excluded to keep the review focused. The goal is rather to showcase the approach that researchers have taken toward this problem by going over some of the relevant work.

### 3.3.1 Image-to-Image Translation

Image-to-image (I2I) translation is the process of learning to map from one image to another image (Alotaibi, 2020). This could be a mapping from a healthy image to an image with pathological identifiers. The differences between the input and the output images can then be used to reveal medical knowledge. For this Generative Adverserial Networks (GAN) or Variational Auto Encoders (VAE) are used.

RegGAN has shown to be to most effective I2I solution on medical data (Kong et al., 2021). The problem with I2I, in the medical domain, is that it is difficult to find aligned image pairs in the real world. The authors used magnetic resonance images of the brain, which they augmented with varied levels of noise and synthetic misalignment by scaling and rotating the image. RegGAN outperformed previous state-of-the-art both for aligned and unaligned pairs as well as from no noise to heavy noise.

Within I2I translation there is also work taking advantage of newer architectures, such as the Transformer. The Swin transformer-based GAN showed promising results on medical data with experiments outperforming RegGAN on the same dataset (Yan et al., 2022).

### 3.3.2 Data Mining Techniques

Data mining is the process of extracting knowledge from big data by finding relevant patterns and relationships (Neha & Vidyavathi, 2015). One such technique is clustering, which is central to this thesis. This has been used in various aspects of knowledge discovery in the medical domain but rarely directly in medical imaging.

Erro et al., 2013 found that K-Means clustering could be used to find subgroups of patients that had not yet been treated. Four new unique subgroups were revealed. This showed promising results for K-Means clustering but the research did not use images. Dy et al., 2003 proposed a system for medical image retrieval by first searching for the image given a major class (known class) and then by the learned identifiers that were not previously labeled. They showed that using clustering, unlabeled subclasses could be revealed and used to find similar images. This was useful as it increased the accuracy of a doctor's diagnosis from 30% to 63%.

Other work looks at how data mining techniques can be used not to directly provide new medical knowledge by AI but rather to use AI to provide a user with better information such that the user can extract new medical knowledge. Schultz and Kindlmann, 2013 proposed a visual solution for practitioners that uses spectral clustering at its base to provide information about 2D and 3D medical data. Although they make use of spectral clustering they acknowledge that there is no clustering method that is best in all cases.

### 3.3.3 Explainable AI

Research on XAI in the medical domain is dominated by usage for ethical and legal reasons, to increase trust and privacy, or to expose bias in models (Sheu & Pardeshi, 2022). The use of XAI for medical knowledge discovery is rarer although some see it as an area with great potential (Nagahisarchoghaei et al., 2023).

Ghorbani et al., 2019 showed that one could cluster images, give the groups of images importance scores, and thereby get explanations as to what parts are important across a whole class. The method uses super-pixel segmentation to cut the image into smaller chunks. This is done for all images in a class. The segments are then clustered and the importance of each segment group is calculated. This results in explanations that contain information about the features that are of importance across each class as a whole. The authors used a general dataset to evaluate their method but it does not seem to be a stretch to apply it to the medical domain. In such a case, this work could potentially reveal medical knowledge by grouping types of identifiers. This is similar to the goals of this thesis, however, using a different method. In this method, explanations are the result. This stands in contrast to the work done in this thesis where explanations are an integrated part of the method that works to improve the grouping.

Hicks et al., 2021 showed that one could leverage XAI for medical knowledge discovery. They used a ResNet to automatically analyze ECG data. With their model, they could predict the sec of the subject with an accuracy of 86%, something they say is nearly impossible for a human cardiologist. To reveal this medical knowledge learned by the model, the authors used XAI. They chose to modify Grad-CAM, the method detailed in Section 3.1.2, to work on the plotted ECG data. This gave a visual explanation of the parts the model finds important for the prediction and revealed what the authors called "new insight into electrophysiology". Although this paper does not use images, like the work in this thesis, it does use 2D visual explanations to reveal new medical knowledge. No papers were found to directly use the explanations of image classifiers in a similar manner to the work in this thesis.

# Chapter 4

# Methodology

Medical institutions sit on a large amount of electronic medical data. This data has historically been underutilized (Liao et al., 2010). Recent literature on the topic, however, has shown that deep learning is a promising direction for making use of this data. This data may be labeled in a rudimentary way, such as healthy and not healthy. The challenge targeted in this thesis is, therefore, to find the characteristics that provide a deeper understanding of the characteristics of the different classes, such as diseased and healthy. In this chapter, a new methodology is proposed to address this problem.

## 4.1 Overview

Suppose that we have a deep learning method that automatically classifies individuals as either healthy or sick. We can imagine that individuals can be classified as sick for many different reasons. For instance, the gastrointestinal tract can consist of several different malformations, all being within the broader class of sick. In this thesis, we suggest a method that will use a combination of DL, XAI, and clustering to get further insight into the properties of the broader classes such as sick and healthy.

An overview of the method developed and implemented in this thesis can be seen in Figure 4.1. The green border represents healthy data while the red with a dotted border represents the flow of pathological data. The flow shown in the figure is for revealing subclasses of the pathological data, but the method could be used for the same purpose on healthy data as well. The method follows four main steps with each step summarized below.

**Step 1** Train a black-box image classifier on healthy and pathological data (binary classification). The data is detailed in Section 4.2 and Section 4.3 while Section 4.4 presents the selected black-box classifier.

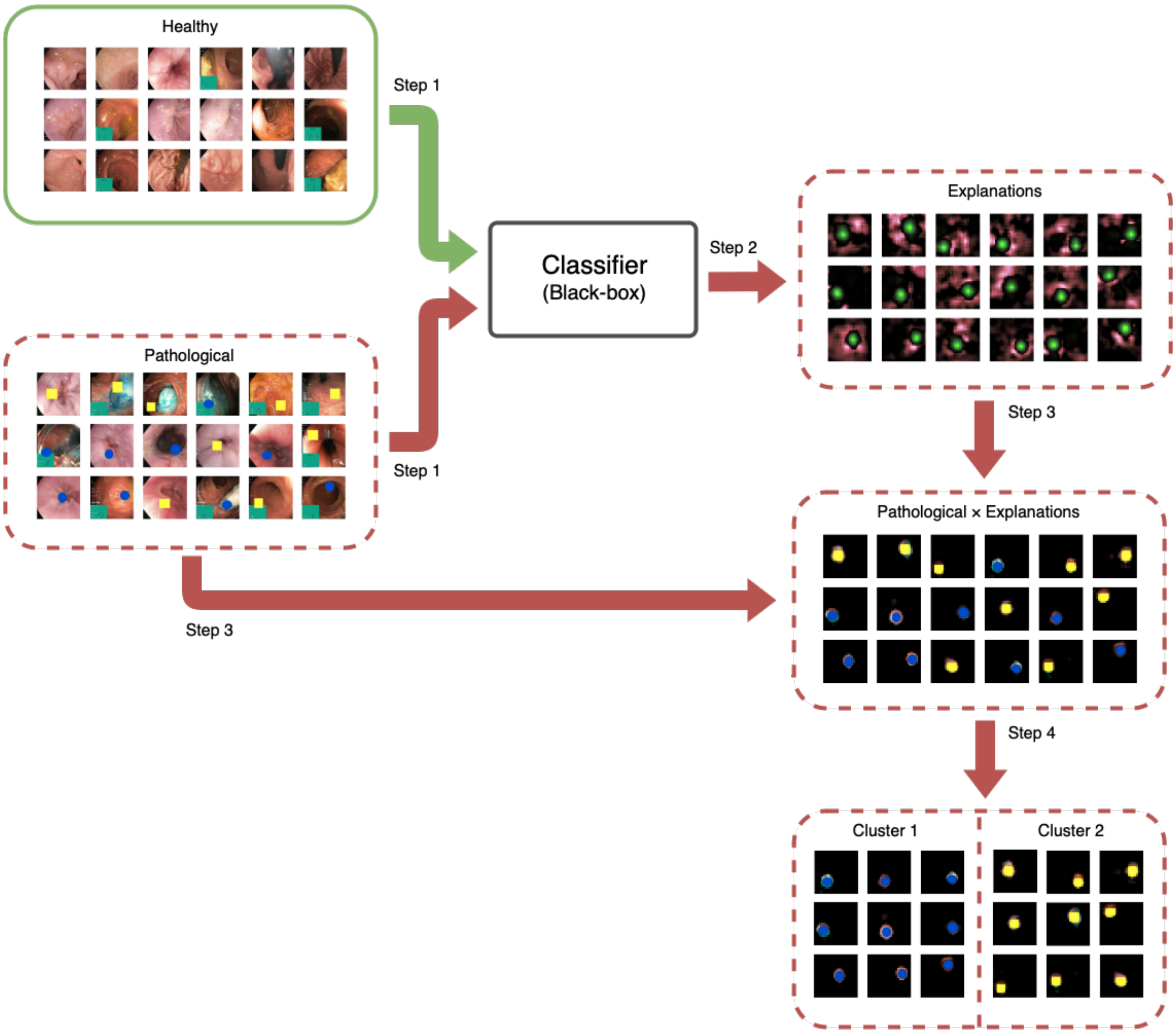

Figure 4.1: Overview showing the four main steps of the proposed method.

**Step 2** Extract explanations for each prediction on the pathological data using XAI techniques as set out in Section 4.5.

**Step 3** Create a final version of the images by using the explanations as a mask applied to the pathological images to highlight areas of importance for the classifier to separate between the two classes and reduce the noise.

**Step 4** Steps 2 and 3 identify important characteristics when predicting the sick images. However, to further get a better overview of the characteristics of subclasses of the pathological data, we suggest clustering the images masked by their respective explanation. We refer to this as Explanation-masked Clustering (XMASK Clustering), which will be described in Section 4.6.

We evaluate this method using two sets of experiments. Each experiment compares the proposed method with its traditional counterpart (no explanations). Section 4.7

contains experiment-specific details.

## 4.2 Data

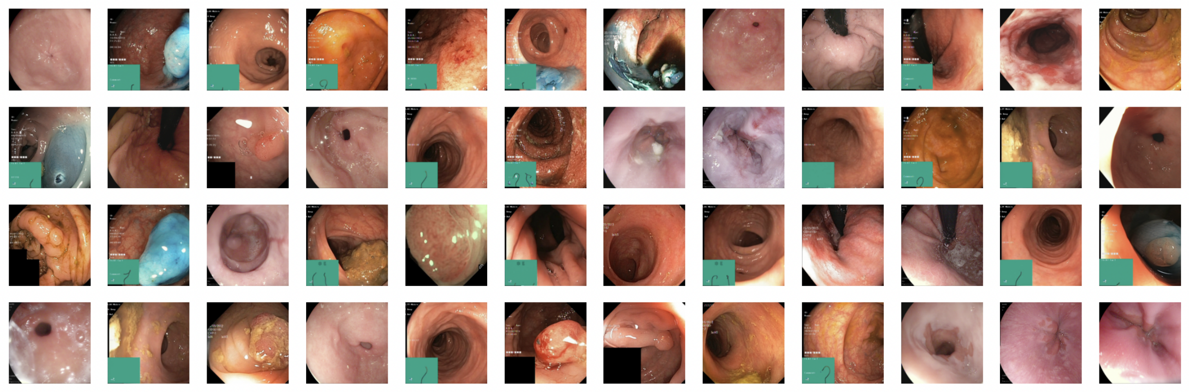

Figure 4.2: A sample from the HyperKvasir dataset showing images from the upper- and lower GI tract.

The data source used as the base for all experiments in this work is the HyperKvasir dataset compiled by SimulaMet (Borgli et al., 2020). This is a large image dataset of the gastrointestinal tract taken from gastro- and colonoscopy examinations performed at Bærum Hospital in Norway. The images were taken using a Pentax colonoscope (Pentax Medical Europe, Germany). Some images contain extra information in the form of a picture-in-picture located in the bottom left corner recognizable by its distinct green background. These are images taken by an Olympus ScopeGuide™, a device used to image the colon (Olympus Europe, Germany).

The dataset contains 10,662 labeled and 99,417 unlabeled images where every label is reviewed by more than one expert in the field and is therefore assumed to be highly accurate. The dataset also includes videos containing 889,372 video frames. This is a large amount of medical data from an unknown amount of patients. Medical data requires special care when it comes to privacy (Price & Cohen, 2019). The Norwegian Data Protection Authority (Datatilsynet) deemed this dataset fully anonymous which allowed it to be openly available. No extraordinary security or privacy measures have therefore been taken when working with this dataset. Figure 4.2 shows a sample from this dataset.

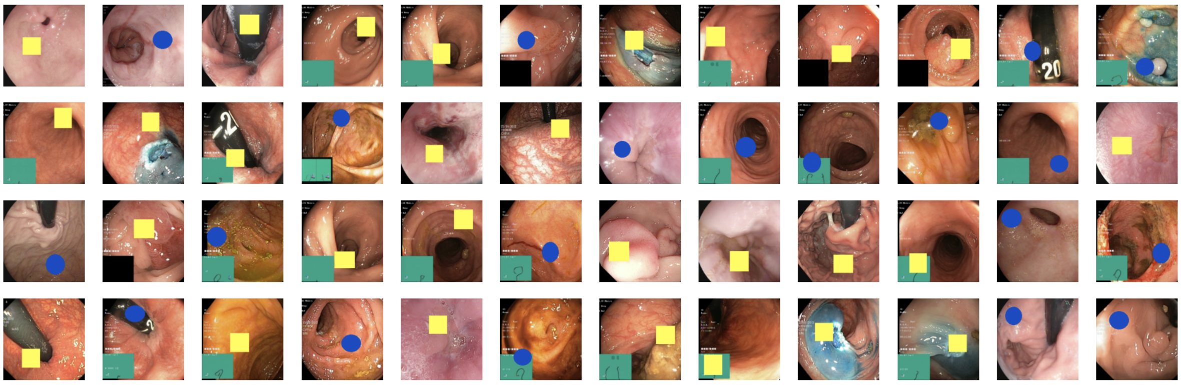

Figure 4.3: Pseudo-real data samples.

## 4.3 Pseudo-Real Data

Pseudo-real datasets created using the HyperKvasir dataset as its base are used in all experiments. The datasets are partially synthetic meaning some samples are real while others are real with a synthetic overlay. Images labeled as healthy are kept as is while those representing pathological findings are pseudo-real in that the real image is used as the base with a colored shape applied on top of the image. The pseudo-real images use the same images as those in the healthy dataset, thus only differing in the colored shapes applied on top. Figure 4.3 shows what this can look like with yellow rectangles and blue ellipses as synthetic pathological identifiers. Applying colored shapes to represent a pathological identifier serves three main purposes. Firstly, it increases the differences between healthy and pathological images. This is important as it allows a classifier trained on the data to learn the differences between the classes, a problem that becomes increasingly difficult with more classes and obscure differences. A classifier that can make out these differences is required for the experiments on explanation-based knowledge discovery detailed in Section 4.7 which is the core focus of this thesis. Secondly, since we know the true explanation, given by the presence or absence of the colored shapes, it allows evaluation of a model's performance, in terms of accuracy or in terms of explanations of predictions, to be evaluated by those who do not have sufficient medical expertise. Lastly, it allows for customization and opens for a wider range of experiments. With the number of possible synthetic classes only being limited by the combinations of colors and shapes, the researcher is given the power to make a problem as simple or complex as needed. Commonly, healthcare-related data, such as this, is unbalanced. The pseudo-real dataset allows us to dynamically change the data to be as balanced or unbalanced as one wishes and enables experiments that the original dataset could not.

## 4.4 Classification

The image classification part of this method is done using a black-box image classifier that solves the binary classification problem of differentiating healthy and pathological images. The purpose of the classifier is to automatically learn the characteristics that define these classes. This is leveraged in later steps. To do this a model appropriate for the task must be chosen. The ResNet architecture with 50 layers was selected for the experiments done in this work. Table 4.1 provides details on the architecture. All input images are scaled to $224 \times 224$ pixels with all three RGB color channels being used. Further, the model architecture follows that of ResNet50 until the classification head. The classification head is task-specific and is replaced with a single neuron giving a prediction in the form $\hat{y} \in \mathbb{R} \mid 0 \leq \hat{y} \leq 1$ after softmax is applied. This architecture was chosen as it has shown good results on classification tasks within medical imaging. It is also a model that previous research on XAI has used to extract explanations and can therefore be seen as a safe base model to select for the experiments in this work. The proposed method does, however, not require a specific architecture.

| |
|---|
| Input $x \in \mathbb{R}^{224 \times 224 \times 3}$ |
| Conv 64, 7×7, stride=2, BN, ReLU |
| Max Pool, 3×3, stride=2 |
| Residual Block × 16 |
| Average Pool, 2×2, stride=2 |
| Fully-connected $1$ |
| Output $\hat{y} \in \mathbb{R} \mid 0 \leq \hat{y} \leq 1$ |

Table 4.1: Classifier architecture overview

## 4.5 Explanations

This section describes the approach used for creating explanations. It is based on the literature review as well as a judgment based on the target case of this thesis as well as the data and selected model. This is necessary as the literature review did not reveal an XAI technique that is universally best.

A multitude of methods for extracting explanations from image classifiers was found, each with its benefits and trade-offs. The occlusion-based method by Zeiler and Fergus, 2014 was selected and built upon to create the explanations used here. An occlusion-based method was selected as they have properties that are wanted when creating explanations for critical applications and trade-offs that are acceptable in the target use case.

Given an image $I \in \mathbb{R}^{W \times H \times 3}$ and a black-box classifier $f(I) \rightarrow \hat{y} \in \mathbb{R} \mid 0 \leq \hat{y} \leq 1$, an explanation is received by altering $I$ and recording changes in $\hat{y}$ for each pixel $j$. The alteration is a 2D patch of pixels $P$ of size $P_W \times P_H$ moved over the image with stride $s$. The patch is applied $N$ times until all pixels have been covered. Each patch is applied to a copy of the original image. The colors of the touched pixels are replaced with gray color values (128 in all color channels for RGB 0-255). As the stride increases, the computation decreases, but with the cost of decreased detail. Gray was selected as this is the color used in most research found on the topic. There is, however, some research using black pixels. The color does not fundamentally change this method. Let this color be represented by the constant $C$. To simplify the calculation, the patch $P$

can be represented with the same dimensions as the image in the first two dimensions with 0's in the position of the patch and 1's in all other positions. This allows for matrix multiplication to be done cleanly. Occluded input using a given patch can then be represented by:

$$g(I, P) = f(I \odot P + (1 - P) \odot C) \tag{4.1}$$

The importance map $Y$ showing the importance of each pixel, $j$, can then be mathematically formulated as the following:

$$Y = \forall j \in I = \frac{1}{|P(j) = 0|} \sum_{i=1}^{N} f(I) - g(I, P_i) \tag{4.2}$$

These raw values do not make good visualizations. Common practice is therefore to normalize between 0 and 1 prior to visualization. Since $Y$ contains both negative and positive values, representing pixels that affect the model's confidence both negatively and positively for the current class, $Y$ is normalized between 0 and 1 for all positive values and 0 and -1 for all negative values. The attribution methods found in the literature review only keep the positive values but in Section 3.1.3 research on visualizations showed that negative values are important for creating complete explanations and are therefore also used in this method. The mathematical formulation of the normalization procedure is left out for brevity. Further, a smoothing function, $S$, is applied to the importance map to remove noise from lesser important pixels and to boost those of higher importance. This step was inspired by (Sundararajan et al., 2019) where they clip the values, creating a hard boundary between important and unimportant pixels. In this work, however, it was found that a smoothing function provided better visualizations. The smoothing function takes three arguments. First, the value to smooth, $x$, second the offset $\theta$, and third the strength $s$. The formula can then be defined as:

$$S(x, \theta, \sigma) = \begin{cases} \frac{x^\sigma}{\theta^{\sigma-1}}, & \text{if } x \leq \theta \\ 1 - \frac{(1-x)^\sigma}{(1-\theta)^{\sigma-1}}, & \text{otherwise} \end{cases} \tag{4.3}$$

The final explanation can then be defined as $S(Y, \theta, \sigma)$ where $\theta$ and $\sigma$ are hyperparameters that need to be selected for each use case. The function is visualized in Figure 4.4 with $\theta = 0.1$ and $\sigma = 8$.

As shown in detail above, this method relies only on altering the input of the model and recording changes in output and can therefore be seen as model-agnostic. This is important for the future applications of this method. It allows for the best model to be selected based on its ability to identify characteristics in medical imaging rather

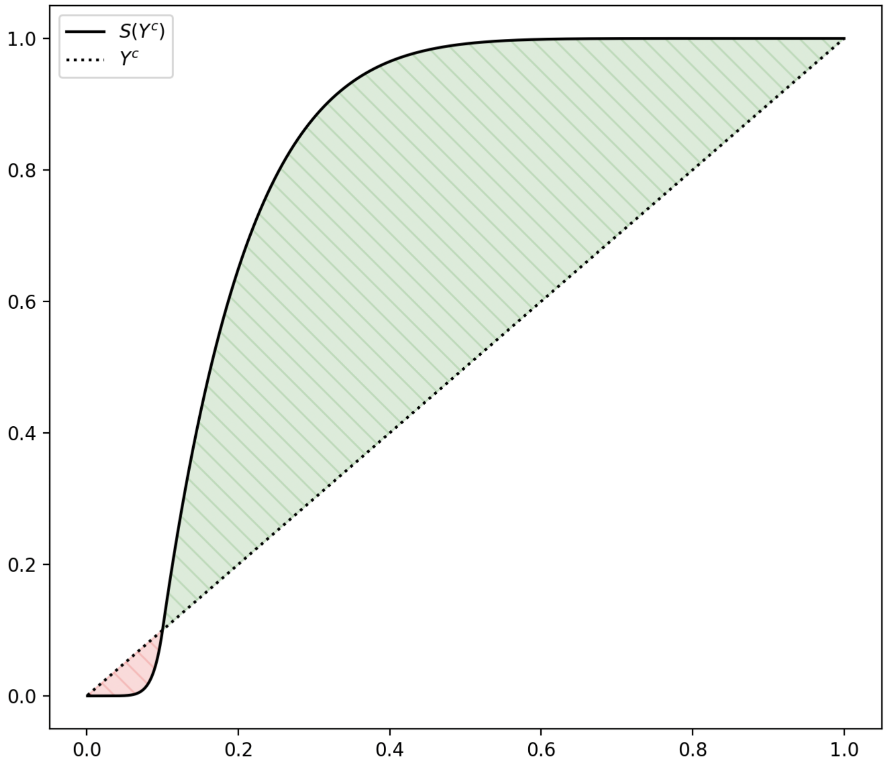

Figure 4.4: Visualization of the smoothing function applied to the explanations.

than based on its compatibility with the XAI method. This stands in contrast to the tight coupling that some gradient-based methods can have to the model architecture. Another reason for using Zeiler and Fergus, 2014 occlusion-based method as a starting point for the method presented here is that it was found to have the highest ease of use in terms of a human's understanding of the explanation (van der Velden et al., 2022). This is critical in the medical domain where various non-XAI experts must make sense of an explanation.

This method does, however, have a downside in terms of computational efficiency. In Section 3.1.1 of the literature review, a host of other occlusion-based methods were reviewed. These mostly aimed to decrease computation time. Significant improvements were made but it came with the cost of decreased detail in the visualizations and in some cases decreased faithfulness to the model's predictions. In this work, higher-detail explanations are considered more important than low computation time.

## 4.6 Explanation-Masked Clustering

Explanation-masked clustering (XMask Clustering) is a novel method of clustering where visual explanations are used to mask images prior to clustering. As the explanations represent the importance of each pixel in an image, using this as a mask results in an image that is simplified by removing unimportant parts. Clustering can then be done on images that only represent characteristics of importance. In this work, clustering is done to reveal groupings of pathological identifiers. The importance map, $Y$, contains information about the importance of each pixel for both the pathological and healthy classes. To get the importance map for only the pathological class the following must therefore be done:

$$Y^{pathological} = \forall j \in Y = \begin{cases} j, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.4}$$

A mask for images in the pathological class can then be created from the explanation for this class. The mask, $M$, is then mathematically formulated as the following:

$$M = S(Y^{pathological}, \theta, \sigma) \tag{4.5}$$

The resulting mask is then a matrix $M \in \mathbb{R}^{W \times H} \mid 0 \leq M \leq 1$ where $W$ and $H$ are equal to the width and height of the original image. The masked image is then received by element-wise multiplication of the mask and the image: $I \odot M$. Direct clustering of the masked images is not practical with 50 thousand features for a $224x224$ image or more than a million for higher definition images such as $1024x1024$. A form of feature extraction is required. This is done by running the masked image through an image classifier and taking the output prior to the classification layer. For the classifier detailed in Section 4.4, the penultimate layer is the average pooling layer outputting a 256-dimensional vector. This vector could be clustered directly, however, 256 dimensions is larger than ideal in terms of computation and clustering accuracy. It is therefore reduced to a lower dimension using Singular Value Decomposition prior to clustering. Clustering is done using K-Means. As K-Means requires the number of clusters, $K$, to be given, this does not work directly with our method as the number of clusters that may exist within the pathological data is unknown. The number of clusters must therefore be found by clustering multiple times with an increasing amount of clusters and evaluating the cluster metrics. The $K$ with the highest Silhouette Coefficient is selected. The Silhouette Coefficient is used as the metric as it gives a direct, measurable, number representing how well-defined the clusters are. The intuition behind this is that well-defined clusters (high Silhouette Coefficient) indicate that the data is correctly clustered as it is difficult to separate a well-defined cluster while poorly defined clusters (low Silhouette Coefficient) indicate that there may exist multiple clusters in the given cluster. This can produce more than one viable solution depending on the data but this is also true for other methods of selecting $K$, such as the Elbow method where the inertia is plotted and the elbow point in the graph represents the correct number of clusters.

With the data clustered, groups of pathological identifiers in the images are revealed. A medical professional would then have access to an explanation as to why the classifier predicted an image as pathological, an image masked with this explanation that highlights the areas of importance, and a grouping based on the characteristics of the pathological data.

## 4.7 Experiments

This section contains details on the experiments with the new methodology. The aim of the experimentation is to evaluate the feasibility and viability of the proposed methodology.

### 4.7.1 Datasets

Two datasets are used for experimentation. The datasets differ in size and content. They do, however, follow the same data preparation procedure.

**Preprocessing** Some preprocessing is required to prepare the dataset for use in machine learning applications. The raw images have an aspect ratio of 4:5. The classifiers used in these experiments, however, use squared images. The images are therefore cropped to a 1:1 aspect ratio. A bottom left crop ensures that the entire Olympus ScopeGuide™ image is included. This does remove 20% from the top of the images and thereby creates some information loss. This is acceptable and aligns with the methods used in the official experiments by the HyperKvasir dataset creators. The images are then scaled to $224x224$ pixels, as expected by the classifier. Finally, the images are normalized between 0 and 1.

**Augmentation** All datasets follow the same data augmentation procedure. This is a standard procedure used to combat overfitting and increase the generalizability of the model. The training data is augmented in the following way: 50% chance of randomly flipping the image horizontally or vertically and uniform randomly rotated up to 90 degrees.

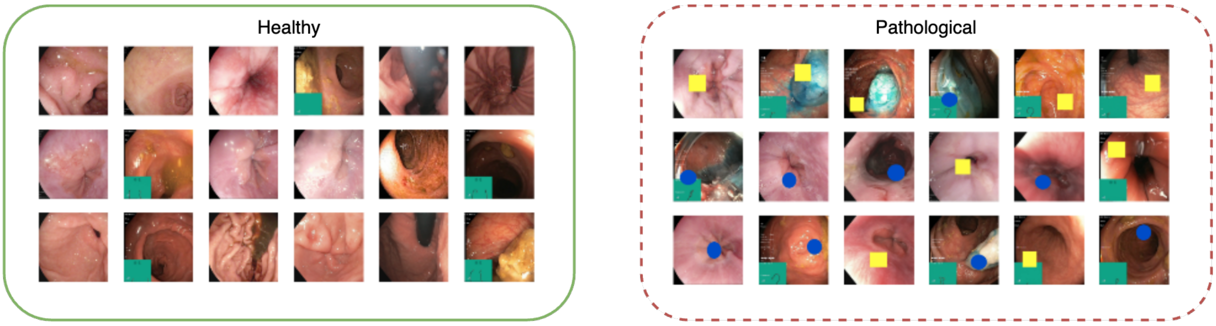

Figure 4.5: Dataset A

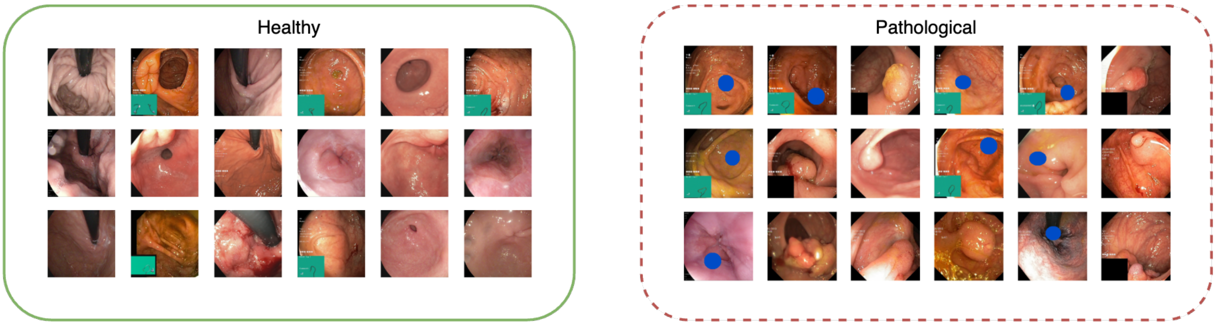

Figure 4.6: Dataset B

**Dataset A** Dataset A, visualized in Figure 4.5, contains 20,000 images evenly split between the healthy and pathological classes. The healthy class uses all 10,000 HyperKvasir images without any edits. The pathological class contains pseudo-real data created by taking the same 10,000 images and applying colored shapes. A single yellow rectangle or blue ellipse, with a width and height randomly set between 20-25%, is added to the image, each at a rate of 50%.

**Dataset B** Dataset B, visualized in Figure 4.6, contains 2,056 images labeled as healthy in the HyperKvasir dataset and 2,056 images with half being real pathological images and half being pseudo-real. The first 1,028 images of the pathological set are real images of polyps. The other 1,028 are healthy images with a blue ellipse, representing another pathological identifier, added in a similar manner as Dataset A. The blue ellipse was chosen as it has similarities to polyps in terms of size and shape, adding to the challenge of this dataset.

### 4.7.2 Classifiers

The experiments require three classifiers: one general pre-trained classifier and two classifiers trained on their respective datasets. This section covers experiment-specific details on the classifiers not covered in Section 4.4.

**Pre-trained Classifier** The pre-trained classifier is a ResNet50 trained on 1,000 classes of ImageNet (Russakovsky et al., 2015). Model weights are downloaded from Pytorch [1]. This model serves two purposes: (1) being the source model for transfer learning and (2) being a general feature extractor.

**Classifier A & B** Classifier A is trained on dataset A using transfer learning to fine-tune from the ImageNet pre-trained classifier. Classifier B differs only in that it uses dataset B. Both classifiers are trained using the same procedure and only differ in the dataset.

**Training** The goal of the classifier in these experiments is to get a well-performant model for the sake of explanations and feature extraction prior to clustering, not to find the best possible model. Safe and well-established choices have therefore been made. The classifiers are fine-tuned on their respective datasets. The data is split into training, validation, and test sets with 64%, 16%, and 20% of the data respectively. The data is batched with a batch size of 64. Loss is calculated for each batch using Binary Cross Entropy loss function. Stochastic Gradient Descent with a learning rate of 0.001 and a momentum of 0.9 is used. The learning rate is decayed by a multiplicative factor of 0.1 every 7 epochs. With these parameters, the model is trained until early stopping based on no change in the F1 score on the validation set using a patience of 5 epochs. The models are trained on a V100 GPU.

### 4.7.3 Experiment Baseline

The baseline is meant to represent the traditional method of clustering and is the method of comparison. In order to accurately compare the methods, all parts except for XMask Clustering specific elements are kept constant between the baseline and experiments, allowing for precise measurements of any differences. With the baseline method, the images are kept as, this stands in contrast to XMask Clustering where the images are masked based on the explanations for the given image. Features are then extracted from these images using the same feature extractor as with the XMask Clustering experiment. The clustering procedure of these features also matches that of the experiment, further detailed in the respective experiment section.

[1] https://pytorch.org/vision/stable/models.html

### 4.7.4 Experiment A

Experiment A is an experiment with the proposed methodology using dataset A and classifier A. The purpose of this experiment is to evaluate the proposed methodology's ability to reveal subclasses in pseudo-real medical data as compared to the traditional method of clustering (baseline). The baseline differs in that it does not use explanations. The proposed methodology is followed as previously detailed. This requires setting some hyperparameters for the explanations. A patch size of $P_W, P_H = 24$ and a stride $s = 8$ is set. For the smoothing function, $\theta = 0.1$ and $\sigma = 8$ are used. The clustering also requires setting a hyperparameter, the number of dimensions to reduce down to before clustering. For this experiment, 32 dimensions are used.

### 4.7.5 Experiment B

Experiment B uses dataset B and classifier B. This experiment goes beyond pseudo-real by having one real pathological identifier (polyps). This is one step closer to the targeted real-world scenario and allows for a more robust evaluation of the method. The hyperparameters used for the explanations in this experiment are as follows. A patch size of $P_W, P_H = 64$ and a stride $s = 16$ is set. For the smoothing function, $\theta = 0.1$ and $\sigma = 8$ are used. Lastly, the features are reduced to 32 dimensions.

# Chapter 5

# Results

This chapter presents the outcome of this thesis work. The results from the experiments, detailed in Section 4.7, are presented and analyzed by looking at each individual logical step outlined in the proposed methodology.

## 5.1 Overview

The results are presented in three parts. First, results on the training and evaluation of the experiment classifiers. Second, the created explanations and an analysis of their accuracy and limitations. Third, results from clustering with and without the proposed XMask Clustering technique. The results can be summarized as the following:

1. **Classifier Results** The black-box classifiers both learned to classify their respective datasets with a perfect score on the test set.
2. **Explanation Results** The explanations visually show that the classifiers correctly focus on the expected areas of the image. Experimentation also showed that the best results are highly reliant on the selected hyperparameter values.
3. **Clustering Results** Experiments with the proposed XMask Clustering technique showed that it results in higher-quality clusters. The clustering correctly groups subclasses of the pathological data and with that reveals medical knowledge.

## 5.2 Classifiers

The classifiers detailed in Section 4.7.2 both achieved a perfect 1.0 F1-score on their respective test datasets. Classifier A must learn to differentiate gastrointestinal images with and without yellow rectangles and blue ellipses. The added noise from the base images on the pseudo-real data makes this a more difficult problem than with fully synthetic data but the ResNet50 fine-tuned from the pre-trained classifier learns quickly and perfectly classifies all test samples after 3 epochs.

Classifier B, trained on dataset B, requires learning more complex features with one of the two pathological identifiers being real images of polyps and the other being blue ellipses. The polyps are both larger and smaller than the previously applied rectangles and ellipses and have significant variations in shape with colors that are similar to the rest of the colon. This is a more complex challenge than dataset A that the classifier manages to overcome. Classifier B reaches a perfect 1.0 F1-score after 8 epochs. Training takes longer than classifier A, in epochs, likely due to dataset B being more complex and one-fifth of the size. The end result is the same across both classifiers as they successfully classify the test data. It can, therefore, be assumed that the models have extracted knowledge from the datasets. These classifiers are thus fit for further use in creating explanation masks and for extracting features for clustering.

## 5.3 Explanations

Explanations visualize what the models see and focus on for a prediction. Creating these explanations is step two in the methodology, seen in Figure 4.1. As the method only uses explanations for the pathological data, the resulting explanations shown here only focus on this class.

An important question of interest for the explanations is whether they show a focus on the expected areas of the images. A visual analysis of images overlaid with their respective explanations was done to answer this question. The experiment classifiers A and B, detailed in Section 4.7.2, were used for extracting explanations for images in datasets A and B. The results are here presented using two samples of typical cases that display the observed findings well.

The first case is that of the pseudo-real data. Figure 5.1 shows two images with the left image showing a pseudo-real sample image with a blue ellipse representing a pathological identifier. The image on the right shows the explanation overlaid on the image. Green areas represent areas that affect the model's prediction of the pathological class positively while red areas represent a negative effect. Higher brightness and visibility of the colors represent higher importance. The explanation shows that there is a large focus on the expected area (the blue ellipse). This shows that the classifier has correctly learned the identifiers of the pathological class. This remains true across a visual analysis of a larger number of samples. The insight that the explanations provide reveals no obvious bias in the classifier. A medical

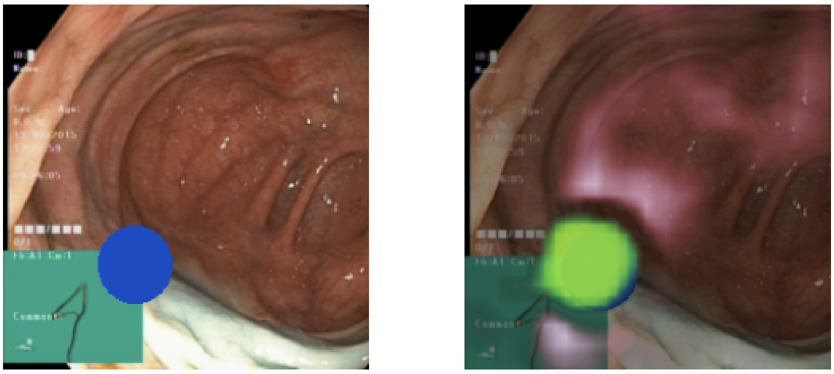

Figure 5.1: Left: A pseudo-real data sample. Right: The sample overlaid with its explanation.

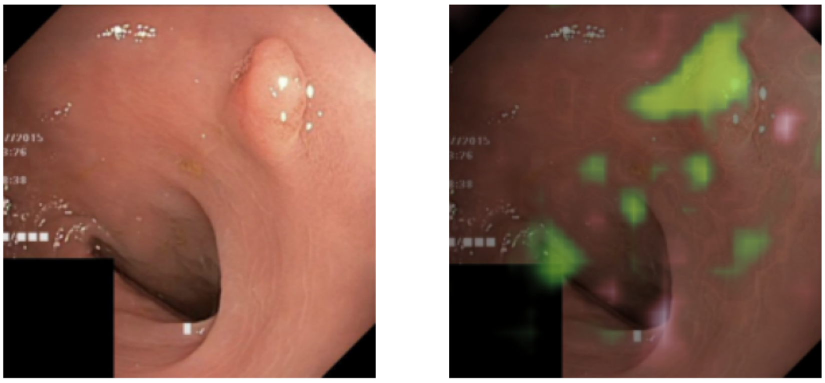

Figure 5.2: Left: A real data sample of a polyp. Right: The sample overlaid with its explanation.

professional utilizing this method could with this get an appropriate amount of trust in the model. This is something the classification F1-score itself cannot provide. The explanations can therefore be said to be a useful tool.

The second case is real medical imaging data. Figure 5.2 shows an image of a polyp (the small growth in the upper right corner of the image) with and without the explanation overlaid. The explanation shows the strongest and largest focus area on the upper part of the polyp. The base of the polyp gets little attribution for the classification. In contrast to the explanation for the pseudo-real sample, this sample shows attribution areas in multiple places. These areas are less bright indicating less importance. No stance can be taken on whether they are relevant from a medical diagnosis perspective as this would require a medical professional.

Although the focus is on the right area, the samples in Figure 5.1 and 5.2 also show

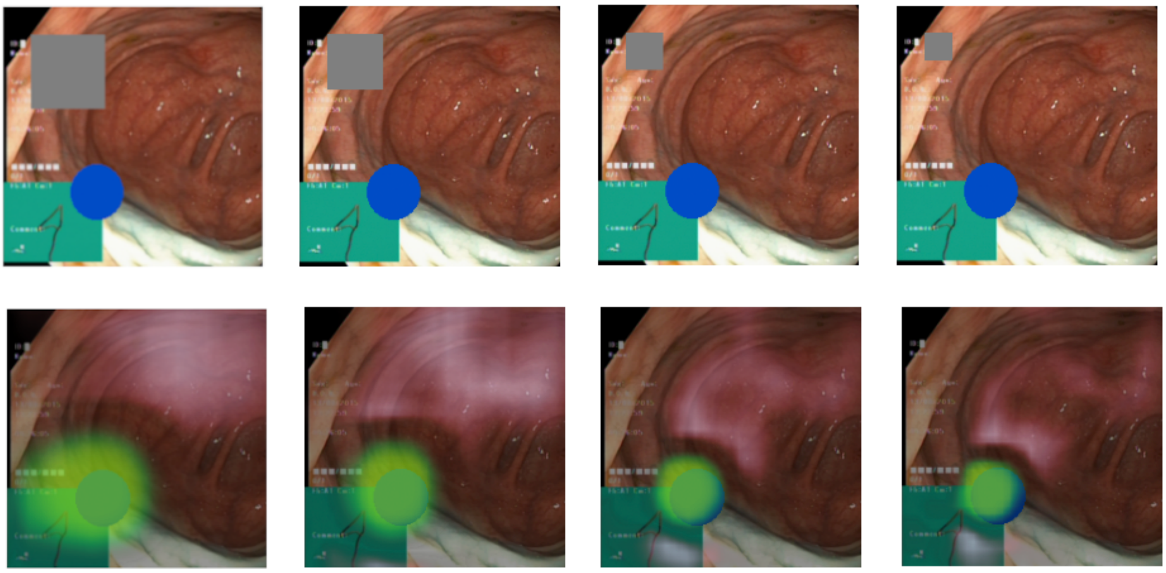

Figure 5.3: Visualization of how the patch size affects the explanation.

that it is slightly offset towards the top left of the pathological identifier. This is likely due to a combination of the selected attribution method, applying the gray rectangles from top-left to bottom-right, the selected patch size, and the classifier architecture. This can be deduced from the fact that this offset is present in all classes and across explanations from both classifiers A and B. Further research would be required to narrow down the source of the offset. How it aligns with the expected focus and the amount of detail is also dependent on the hyperparameters used. Figure 5.3 shows the correlation between the patch size and the accuracy of the explanation. The top row of images shows the image with the gray patch for size reference. The bottom row shows the corresponding explanations. The figure starts on the left with a patch size of 64 pixels, then 48, 32, and lastly 24 pixels. The first patch is roughly 150% the size of the ellipse and causes significant attribution on areas outside of the ellipse. The next patch is roughly the same size as the ellipse and still causes large areas outside of the ellipse to give attribution, although less than the previous patch. Reducing the patch to 75% of the size of the ellipse gives a more accurate attribution area. Further reducing the patch size to 50% of the size of the ellipse further reduces the amount of attribution outside of the ellipse but also removes some correct attribution from the ellipse itself. From this one can see that the explanations are highly reliant on the selected hyperparameter for patch size. One can say that as the patch size increases the recall increases and as the patch size decreases the precision increases. There is therefore a trade-off that must be made. To ensure that all parts of the pathological identifier are in focus, it is advisable to choose a larger patch size. On the other hand, if you want to avoid inaccuracies in the explanation, it is better to choose a smaller patch size.

The explanations showing the areas of importance are not only for detecting bias or insight into the model for increased trust but also for medical knowledge discovery. As the data is simply labeled as pathological, the explanations showing the model's learned medical knowledge can be an aid in knowledge discovery through the medium of a visual explanation. The explanations are further used for more direct medical knowledge discovery in the XMask Clustering experiments detailed in Section 5.4.

## 5.4 Explanation-masked Clustering Experiments

In Section 4.7 two experiments were outlined. These experiments follow the proposed methodology of XMask Clustering for two different datasets using two different classifiers. In this section, the experiments are presented chronologically, beginning with Experiment A and then proceeding to Experiment B. Each experiment is run twice, once using the experiment classifier for feature extraction and once using the pre-trained classifier. This is to assess whether using masked input images would yield better or worse results for feature extraction. Additionally, the reliance on the experiment classifier for feature extraction can be determined.

### 5.4.1 Experiment A

This experiment, detailed in Section 4.7.4, evaluates the proposed XMask Clustering methodology for pseudo-real data. A few samples from Dataset A used here can be seen with and without the explanation masks in Figure 5.4. The figure shows that the masks have accurately removed the areas surrounding the synthetic pathological identifiers. Whether this helps in clustering is evaluated using the following two experiments.

The first experiment clusters features extracted using classifier A. Table 5.1 shows the results of clustering with the baseline technique and using XMask Clustering. Both found two clusters through X-Means. The clustering is evaluated on classification ability (assigning a data point to the correct cluster) using a test set that is 20% of dataset A and has never previously been seen by the cluster. Rand index is used to measure the accuracy of the clustering. The classification test results show no change in classification ability with XMask Clustering as both methods correctly classified all data points in the test set. The next metrics used to evaluate the method are the cluster quality metrics inertia, Silhouette Coefficient, and Davies-Bouldin Index. The metrics show a 7.4% decrease in inertia, a 13% increase in Silhouette Coefficient, and a 15.6%

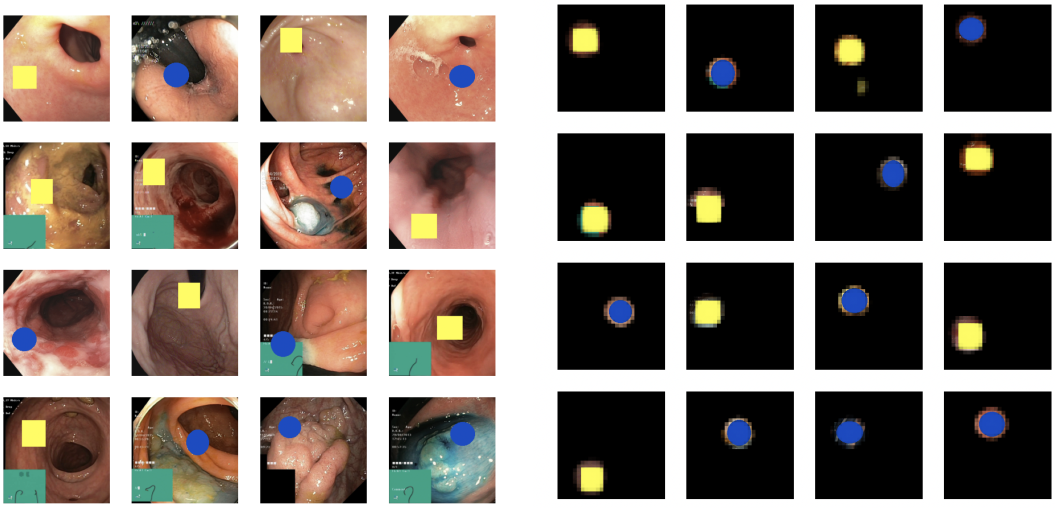

Figure 5.4: Unmasked (left) and masked (right) image samples from dataset A.

| **Methodology** | **Classification** | **Cluster Quality** |
|---|---|---|
| | (Rand index) | (Inertia / Silhouette / Davies-Bouldin) |
| Baseline | 1.0 | 43,741 / 0.484 / 0.858 |
| XMask | 1.0 | **40,485** / **0.548** / **0.724** |
| *Improvement* | *±0.0* | *-3,256 / +0.064 / -0.134* |

Table 5.1: Clustering results for experiment A using classifier A as the feature extractor. The first column represents the methodology that is used. The second column represents the results of a classification test using samples previously not seen by the cluster. The third column shows cluster quality metrics. Bold indicates the best values.

decrease in Davies-Bouldin Index. With this, it is safe to say that the clustering quality is improved across all metrics when using XMask Clustering.

The second experiment uses the general pre-trained classifier as a feature extractor. This experiment aims to show the method's sensitivity to changes to the feature extractor. Table 5.2 shows the results of the clustering. X-Means resulted in two clusters for both although more distinctly for the XMask Clustering method. The baseline method, clustering unmasked images, had close to a random result on the classification test with a Rand Index of 0.498 meaning that it could not cluster the features in any meaningful way. The XMask Clustering method, however, managed a 0.995 Rand Index. This is likely because the masked images are significantly simpler and thus the method is less sensitive to changes in the feature extractor.

| **Methodology** | **Classification** | **Cluster Quality** |
|---|---|---|
| | (Rand index) | (Inertia / Silhouette / Davies-Bouldin) |
| Baseline | 0.498 | 82,819 / 0.152 / 2.181 |
| XMask | **0.995** | **13,715** / **0.485** / **0.840** |
| *Improvement* | *+0.497* | *-69,104 / +0.333 / -1.341* |

Table 5.2: Clustering results for experiment A using the general feature extractor. The first column represents the methodology that is used. The second column represents the results of a classification test using samples previously not seen by the cluster. The third column shows cluster quality metrics. Bold indicates the best values.

### 5.4.2 Experiment B

This experiment, detailed in 4.7.5, evaluates the proposed XMask Clustering methodology for a combination of pseudo-real data and real data from the HyperKvasir dataset. Figure 5.5 shows samples from the dataset used for clustering. It shows that it can be difficult to correctly isolate the pathological identifiers when they have significant variations in shape. These experiments aim to see whether XMask Clustering also shows benefits in such cases. It follows the same procedure as experiment A with two sub-experiments. These experiments also matched experiment A in that X-Means correctly resulted in two clusters (one cluster for polyps and one cluster for the synthetic identifiers).

Table 5.3 shows the results of the first experiment. Using classifier B as the feature extractor both the baseline and XMask methods get a high score on the classification test with the XMask method correctly clustering all images getting a Rand Index of 1.0 compared to the baseline method's 0.967. The XMask method also created higher-quality clusters with a 12.9% decrease in inertia, a 15.5% increase in Silhouette Coefficient, and a 15.6% decrease in Davies-Bouldin Index. These are significant improvements in all measured clustering quality metrics. This shows that although the masking visually is not perfect it has an improvement on the clustering. There is also no information loss in terms of medical knowledge discovery as all of the clustered, masked, images can be linked to an unmasked image.

The result of the second experiment using the general feature extractor is presented in Table 5.4. The XMask method showed a significant improvement on the classification test compared to the baseline. Similarly to experiment A, XMask Clustering seems to be less sensitive to the feature extractor.

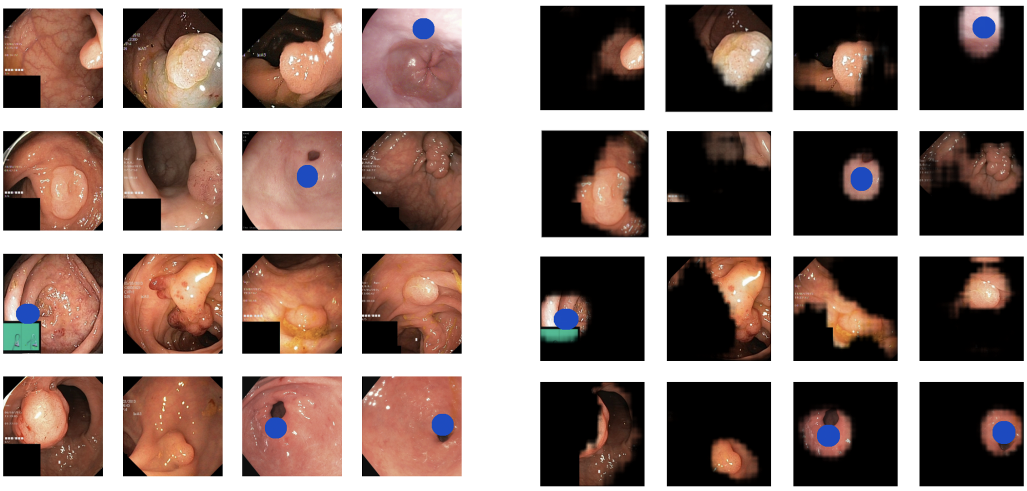

Figure 5.5: Unmasked (left) and masked (right) image samples from dataset B.

| **Methodology** | **Classification** | **Cluster Quality** |
|---|---|---|
| | (Rand index) | (Inertia / Silhouette / Davies-Bouldin) |
| Baseline | 0.967 | 59,017 / 0.373 / 1.124 |
| XMask | 1.0 | **51,406** / **0.431** / **0.979** |
| *Improvement* | *+0.033* | *-7,611 / +0.058 / -0.145* |

Table 5.3: Clustering results for experiment B using classifier B as the feature extractor. The first column represents the tested methodology. The second column represents the results of a classification test using samples previously not seen by the cluster. The third column shows cluster quality metrics.

| **Methodology** | **Classification** | **Cluster Quality** |
|---|---|---|
| | (Rand index) | (Inertia / Silhouette / Davies-Bouldin) |
| Baseline | 0.702 | 75,746 / 0.197 / 1.850 |
| XMask | **0.919** | **55,312** / **0.259** / **1.567** |
| *Improvement* | *+0.217* | *-20,434 / +0.062 / -0.283* |

Table 5.4: Clustering results for experiment B using the general feature extractor. The first column represents the tested methodology. The second column represents the results of a classification test using samples previously not seen by the cluster. The third column shows cluster quality metrics.

# Chapter 6

# Discussion

A new methodology called XMask Clustering was proposed in this thesis. The main contribution is the integration of XAI into the workflow and the use of the relationship between explanations and a deep learning model's knowledge to mask images such that only areas of importance are clustered. This resulted in medical knowledge discovery in two of the stages. First, as the result of the visual explanation revealing the learned knowledge of the black-box model. Using pseudo-real data, and overlaying the explanations on top of such data as seen in Figure 5.1, the explanations could be evaluated for correctness without the need for medical experts. An explanation revealing anything beyond the expected, known pathological identifiers, that a model may have learned, could be seen as a mistake in the explanation method or model. For instance, Figure 5.2 shows that there is a focus on the polyp but that this focus also extends slightly off the poly to the left. This could indicate that there is something of importance in that part of the image. Looking at the explanations for the pseudo-real data in Figure 5.1 however shows a similar offset towards the left. Evaluating the explanations on pseudo-real data may therefore be a method of exposing errors in the model or in the explanation as well as increasing the trust in the correctness of the method prior to using it for real medical data. This is a step that can be taken to ensure that new medical knowledge revealed by the explanations is knowledge and not an error. With this, the second research question can be answered by carefully saying that pseudo-real medical data can be useful in the evaluation of explanations. The second stage where medical knowledge is revealed is with the clustering. As the results show high-quality clustering, according to multiple metrics for quality, that correctly reveals subclasses of the pathological data.

Further discussion related to the work is given in this chapter. The chapter starts by covering discussion points on the proposed methodology, followed by ethical considerations that must be made when working with AI in the medical domain, and lastly, motivation and recommendations for future work.

## 6.1 Faithfulness vs Interpretability

An essential aspect of the method proposed in this thesis is the visual explanation. Within the world of visual explanations, it is important to discuss faithfulness and interpretability. The concept of faithfulness in visual explanations relates to the model's capacity to provide an accurate representation of the learned function. It is worth noting that a certain trade-off exists between faithfulness and interpretability. Despite the importance of faithfulness, it is equally essential to ensure that the explanations are clear and easily understandable to enable effective communication. Interpretability can be difficult for raw explanations. The literature review revealed that it is common to do some processing to clean up the explanation for the purpose of improved interpretability. In this work, a smoothing function was therefore applied to the explanation, removing noise and boosting the brightness of important pixels. Sundararajan et al., 2019 argue that humans cannot see colors linearly and therefore boosting values to make them more visible to human eyes is fair as one may have otherwise missed the importance of an area of the image. In their work, the pixels in the top end of attribution were clipped. With this technique, there is a low amount of gradient between the pixels at the top end. In order to allow for more gradient between the attributed pixels, this work opted for a smoothing function that boosts the values based on a function instead. A hyperparameter can then steer the level of boosting and, thereby, the level of faithfulness in the visualization. Striking a balance between faithfulness and interpretability is critical in producing visual explanations that are both accurate and comprehensible.

## 6.2 Strengths and Weaknesses

The research done in this thesis has both strengths and weaknesses. As the use of XAI for medical knowledge discovery is a highly unexplored area of research, a strength of this work is in the novelty of the research. On the research methodology side, the use of pseudo-real data is notable. The results are much more accessible to the wider AI community when using pseudo-real data, as it does not require medical knowledge to understand. This allows for the work to be accurately evaluated and to generalize to other domains. There is, however, also a weakness that comes with this use. Pseudo-real data does not fully show the usefulness of the proposed methodology on real medical images. No medical professionals took part in this research, and therefore evaluations on real medical data were limited. On the side of the proposed method, this study has demonstrated two distinct advantages that come with XMask Clustering over the more traditional approach to clustering. The first is that explanations are built into the method. This forces the benefits of XAI onto the user rather than being an optional addition to other methodologies. The experiments in Section 5.4 also showed that the method could create higher-quality clusters while maintaining the same link to the original images.

## 6.3 Ethical Considerations

When it comes to the use and research of artificial intelligence, particularly black-box deep learning models utilized in this thesis targeting the medical domain, it is imperative to take into account the numerous ethical considerations that must be made. These considerations must tackle issues ranging from bias and transparency in the system to the protection of patient privacy in the data to the potential societal impact.

One of the main concerns is the potential for bias in AI algorithms, which could lead to unequal healthcare outcomes for different groups of patients. This is not only important after a model has been deployed for use in healthcare but also in the medical knowledge discovery phase as this propagates to later stages. This must therefore be tackled early. Other ethical considerations include the need for transparency in the development and use of AI systems, the protection of patient privacy and data security, and the responsibility of healthcare providers to ensure that AI technology is used in a safe and effective manner. A part of the XMask Clustering methodology presented in this work includes the direct inclusion of explanations. As presented in Section 5.3, the explanations can provide a healthcare professional or medical researcher with visual explanations that give insight into the models. Previous research has shown that this can expose bias and creates an appropriate level of trust in the AI model (Mahmoudi et al., 2022; Ribeiro et al., 2016).

There is an additional matter that warrants attention, and that is the preservation of patient confidentiality through data privacy protection. Deep learning algorithms necessitate the use of extensive data sets for their training, which can have significant privacy ramifications if the data is not managed properly. In this work, all data is deemed fully anonymous by The Norwegian Data Protection Authority (Datatilsynet) (Borgli et al., 2020). The privacy implications are, therefore, low.

The societal impact of the work must also be considered. This research is done with the aim of having a positive societal impact. It is assessed that the downsides of using black-box models in the medical domain have been reduced with the help of eXplainable AI and that the potential for positive impact in medical diagnosis and knowledge discovery outweigh the unfortunate downsides. To ensure a positive societal impact of this work and future work, it is vital to carefully consider these ethical issues as we continue to explore the potential of AI in the medical domain.

## 6.4 Future Work

The importance of the continued work on bringing XAI to the medical domain is a view that is shared by a multitude of researchers on the subject (van der Velden et al., 2022). Some of the motivations for this are as follows. First, explanations can reduce mistakes in medical diagnoses done on the basis of, or with the aid of, a deep learning model's predictions (Hicks et al., 2021). Second, medical knowledge revealed by machine learning techniques may be difficult for a medical professional to trust and use without explanations (Hicks et al., 2021). One should therefore consider always including XAI in research on medical knowledge discovery. Lastly, the potential for improved medical diagnosis either through new medical knowledge or through improved medical care by increased efficiency of diagnoses (Monroe et al., 2021). With these motivations in mind, a few areas of future work related to this thesis is proposed.

In Section 6.2, one of the weaknesses identified in this work was related to the data. The first continuation of this work in relation to data is to increase the number of classes in the pseud-real data. The results in Section 5.4 showed that the proposed method improved clustering and knowledge discovery for two classes but in the real world there may be a higher number of classes and the amount of data in each class may be unbalanced. It would be interesting to see whether the method upholds the good results in such cases. To get closer to a real-world scenario, an area of development that could be explored in the future is the integration of more real data with multiple classes and the inclusion of medical professionals in the evaluation of results. This approach would ensure that the findings are accurate, reliable, and relevant to medical research and the healthcare industry. By incorporating expert insights and knowledge, we can enhance the quality of the data and make more informed decisions. This is a promising direction for the future of healthcare research and innovation. Another avenue that could be explored further is the data format. In this work, 2D image data was used, but medical imaging, such as MRI, can often be 3D data. Some work on 3D MRI data reduces such data down to two dimensions prior to using XAI (Monroe et al., 2021). It would be an interesting future approach to see whether the proposed methodology could leverage such data and find more complex pathological identifiers.

Other areas could also be of interest for future work. It would be interesting to conduct additional experiments to explore the use of deep learning in clustering. Experimentation with XMask Clustering using state-of-the-art Deep Clustering techniques would be exciting and have the potential for more significant improvements. Although the experiments here showed that XMask Clustering was less sensitive to changes in the feature extraction process than the baseline, it is still a significant factor in clustering and could likely benefit from refinement.

# Chapter 7

# Conclusion

The research conducted here aimed to explore the role of eXplainable AI (XAI) and deep learning in medical knowledge discovery. This is done through the following research questions:

1. *Can explanations of a black-box model be leveraged to reveal new medical knowledge while addressing the problems hindering the adoption of deep learning in the medical domain?*
2. *Is using pseudo-real medical data, adding a synthetic layer on top of real medical images, useful for the evaluation of such a technique?*

In support of research question 1, we showed that explanations of black-box models combined with clustering effectively reveal medical knowledge. We demonstrated this through a novel methodology that we called XMask Clustering. This results from a review of relevant literature and experimentation with the new methodology using a high-quality medical dataset. The experiments showed that visual explanations of the black-box model are effective as masks for the medical images and that clustering such explanation-masked images result in better clusters. These clusters are groups of similar identifiers not found in the data labels but rather extracted from the black-box model's learned medical knowledge. Further, to cover the second part of research question 1, the problems hindering the adoption of deep learning in the medical domain, XAI is baked into the methodology. XMask Clustering has the additional benefit of addressing ethical concerns, such as transparency, bias, and trust, through the deep link with XAI. The result is a more reliable and trustworthy solution.

The experimentation also revealed that explanations accurately and correctly highlight areas of importance in real and pseudo-real gastrointestinal medical images. Using pseudo-real data throughout the experimentation phase revealed that it is helpful for the evaluation of explanations as it can uncover errors that may be difficult to detect in real data. This acts in support of research question 2.

As we move forward, it would be valuable to expand on this study by exploring more complex categories and involving medical professionals in the assessment phase. This would enable us to gain a more comprehensive understanding and provide more accurate insights into our findings from a medical perspective. We hope that future researchers will carry on this work with the aim of making a positive impact.

# References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alotaibi, A. (2020). Deep generative adversarial networks for image-to-image translation: A review. *Symmetry*, *12*(10). https://doi.org/10.3390/sym12101705

Alzubaidi, L., Fadhel, M. A., Al-Shamma, O., Zhang, J., Santamaría, J., Duan, Y., & R. Oleiwi, S. (2020). Towards a better understanding of transfer learning for medical imaging: A case study. *Applied Sciences*, *10*(13). https://doi.org/10.3390/app10134523

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms*, *8*, 1027–1035. https://doi.org/10.1145/1283383.1283494

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82–115. https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012

Borgli, H., Thambawita, V., Smedsrud, P. H., Hicks, S., Jha, D., Eskeland, S. L., Randel, K. R., Pogorelov, K., Lux, M., Nguyen, D. T. D., Johansen, D., Griwodz, C., Stensland, H. K., Garcia-Ceja, E., Schmidt, P. T., Hammer, H. L., Riegler, M. A., Halvorsen, P., & de Lange, T. (2020). HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, *7*(1), 283. https://doi.org/10.1038/s41597-020-00622-y

Brodzicki, A., Piekarski, M., Kucharski, D., Jaworek-Korjakowska, J., & Gorgon, M. (2020). Transfer learning methods as a new approach in computer vision tasks with small datasets. *Foundations of Computing and Decision Sciences*, *45*(3), 179–193. https://doi.org/10.2478/fcds-2020-0010

Cui, Y., Song, Y., Sun, C., Howard, A., & Belongie, S. (2018). Large scale fine-grained categorization and domain-specific transfer learning. *2018 IEEE/CVF*

*Conference on Computer Vision and Pattern Recognition*, 4109–4118. https://doi.org/10.1109/CVPR.2018.00432

Das, K., & Behera, R. N. (2017). A survey on machine learning: Concept, algorithms and applications. *International Journal of Innovative Research in Computer and Communication Engineering*, *5*(2), 1301–1309.

Dong, S., Wang, P., & Abbas, K. (2021). A survey on deep learning and its applications. *Computer Science Review*, *40*, 100379. https://doi.org/10.1016/j.cosrev.2021.100379

Dubey, S. R., Singh, S. K., & Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, *503*, 92–108. https://doi.org/10.1016/j.neucom.2022.06.111

Dy, J., Brodley, C., Kak, A., Broderick, L., & Aisen, A. (2003). Unsupervised feature selection applied to content-based retrieval of lung images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(3), 373–378. https://doi.org/10.1109/TPAMI.2003.1182100

Ehrhardt, J., Kepp, T., Handels, H., & Uzunova, H. (2019). Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders, 36. https://doi.org/10.1117/12.2511964

Erro, R., Vitale, C., Amboni, M., Picillo, M., Moccia, M., Longo, K., Santangelo, G., De Rosa, A., Allocca, R., Giordano, F., Orefice, G., De Michele, G., Santoro, L., Pellecchia, M. T., & Barone, P. (2013). The heterogeneity of early parkinson's disease: A cluster analysis on newly diagnosed untreated patients. *PLOS ONE*, *8*(8), 1–8. https://doi.org/10.1371/journal.pone.0070244

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S., & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, *2*(3), 267–279. https://doi.org/10.1109/TETC.2014.2330519

Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Garbin, C., Zhu, X., & Marques, O. (2020). Dropout vs. batch normalization: An empirical study of their impact to deep learning. *Multimedia Tools and Applications*, *79*(19), 12777–12815. https://doi.org/10.1007/s11042-019-08453-9

Gecer, B., Aksoy, S., Mercan, E., Shapiro, L. G., Weaver, D. L., & Elmore, J. G. (2018). Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. *Pattern Recognition*, *84*, 345–356. https://doi.org/https://doi.org/10.1016/j.patcog.2018.07.022

Ghorbani, A., Wexler, J., Zou, J. Y., & Kim, B. (2019). Towards automatic concept-based explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/77d2afcb31f6493e350fca61764efb9a-Paper.pdf

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249–256). PMLR. https://proceedings.mlr.press/v9/glorot10a.html

Godbole, V., Dahl, G. E., Gilmer, J., Shallue, C. J., & Nado, Z. (2023). Deep learning tuning playbook [Version 1.0]. http://github.com/google/tuning_playbook

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1026–1034. https://doi.org/10.1109/ICCV.2015.123

Hicks, S. A., Isaksen, J. L., Thambawita, V., Ghouse, J., Ahlberg, G., Linneberg, A., Grarup, N., Strümke, I., Ellervik, C., Olesen, M. S., Hansen, T., Graff, C., Holstein-Rathlou, N.-H., Halvorsen, P., Maleckar, M. M., Riegler, M. A., & Kanters, J. K. (2021). Explaining deep neural networks for knowledge discovery in electrocardiogram analysis. *Scientific Reports*, *11*(1), 10949. https://doi.org/10.1038/s41598-021-90285-5

Hoffman, R., Miller, T., Mueller, S. T., Klein, G., & Clancey, W. J. (2018). Explaining explanation, part 4: A deep dive on deep nets. *IEEE Intelligent Systems*, *33*(3), 87–95. https://doi.org/10.1109/MIS.2018.033001421

Kapishnikov, A., Venugopalan, S., Avci, B., Wedin, B., Terry, M., & Bolukbasi, T. (2021). Guided integrated gradients: An adaptive path method for removing noise. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5048–5056. https://doi.org/10.1109/CVPR46437.2021.00501

Kong, L., Lian, C., Huang, D., li zhenjiang, z., Hu, Y., & Zhou, Q. (2021). Breaking the dilemma of medical image-to-image translation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (pp. 1964–1978). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2021/file/0f2818101a7ac4b96ceeba38de4b934c-Paper.pdf

LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K.-R. (2012). Efficient BackProp [Series Title: Lecture Notes in Computer Science]. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), *Neural networks: Tricks of the trade* (pp. 9–48). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-35289-8_3

Li, H., Li, J., Guan, X., Liang, B., Lai, Y., & Luo, X. (2019). Research on overfitting of deep learning. *2019 15th International Conference on Computational Intelligence and Security (CIS)*, 78–81. https://doi.org/10.1109/CIS.2019.00025

Liao, K. P., Cai, T., Gainer, V., Goryachev, S., Zeng-treitler, Q., Raychaudhuri, S., Szolovits, P., Churchill, S., Murphy, S., Kohane, I., Karlson, E. W., & Plenge, R. M. (2010). Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care & Research*, *62*(8), 1120–1127. https://doi.org/https://doi.org/10.1002/acr.20184

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. *International Conference on Machine Learning*, *30*, 3.

Mahmoudi, S. A., Stassin, S., Daho, M. E. H., Lessage, X., & Mahmoudi, S. (2022). Explainable deep learning for covid-19 detection using chest x-ray and ct-scan images. In L. Garg, C. Chakraborty, S. Mahmoudi, & V. S. Sohmen (Eds.), *Healthcare informatics for fighting covid-19 and future epidemics* (pp. 311–336). Springer International Publishing. https://doi.org/10.1007/978-3-030-72752-9_16

Mitchell, T. (1997). *Machine learning*. McGraw Hill.

Monroe, W. S., Skidmore, F. M., Odaibo, D. G., & Tanik, M. M. (2021). HihO: Accelerating artificial intelligence interpretability for medical imaging in IoT applications using hierarchical occlusion: Opening the black box. *Neural Computing and Applications*, *33*(11), 6027–6038. https://doi.org/10.1007/s00521-020-05379-4

Moran, M. B., Faria, M. D., Giraldi, G. A., Bastos, L. F., & Conci, A. (2021). Using super-resolution generative adversarial network models and transfer learning to obtain high resolution digital periapical radiographs. *Computers in Biology and Medicine*, *129*, 104139. https://doi.org/https://doi.org/10.1016/j.compbiomed.2020.104139

Nagahisarchoghaei, M., Nur, N., Cummins, L., Nur, N., Karimi, M. M., Nandanwar, S., Bhattacharyya, S., & Rahimi, S. (2023). An empirical survey on explainable ai technologies: Recent trends, use-cases, and categories from technical and application perspectives. *Electronics*, *12*(5). https://doi.org/10.3390/electronics12051092

Neha, D., & Vidyavathi, B. (2015). A survey on applications of data mining using clustering techniques. *International Journal of Computer Applications*, *126*(2).

Ng, A., Jordan, M., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems*. MIT Press. https://proceedings.neurips.cc/paper_files/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf

Papanastasopoulos, Z., Samala, R. K., Chan, H.-P., Hadjiiski, L., Paramagul, C., Helvie, M. A., & Neal, C. H. (2020). Explainable ai for medical imaging: Deep-learning cnn ensemble for classification of estrogen receptor status from breast mri. In H. K. Hahn & M. A. Mazurowski (Eds.), *Medical imaging 2020: Computer-aided diagnosis* (113140Z). SPIE. https://doi.org/10.1117/12.2549298

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. https://doi.org/10.48550/ARXIV.1712.04621

Petsiuk, V., Das, A., & Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. *Proceedings of the British Machine Vision Conference (BMVC)*.

Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine*, *25*(1), 37–43. https://doi.org/10.1038/s41591-018-0272-7

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/https://doi.org/10.1016/j.neunet.2014.09.003

Schultz, T., & Kindlmann, G. L. (2013). Open-box spectral clustering: Applications to medical image analysis. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2100–2108. https://doi.org/10.1109/TVCG.2013.181

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Senaras, C., & Gurcan, M. N. (2018). Deep learning for medical image analysis. *Journal of Pathology Informatics*, *9*(1), 25. https://doi.org/https://doi.org/10.4103/jpi.jpi_27_18

Sheu, R.-K., & Pardeshi, M. S. (2022). A survey on medical explainable ai (xai): Recent progress, explainability approach, human interaction and scoring system. *Sensors*, *22*(20). https://doi.org/10.3390/s22208068

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Shutaywi, M., & Kachouie, N. N. (2021). Silhouette analysis for performance evaluation in machine learning with applications to clustering. *Entropy*, *23*(6). https://doi.org/10.3390/e23060759

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: Removing noise by adding noise.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. A. (2015). Striving for simplicity: The all convolutional net. In Y. Bengio & Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015, workshop track proceedings*. http://arxiv.org/abs/1412.6806

Srinivas, S., & Fleuret, F. (2019). Full-gradient representation for neural network visualization. *Advances in Neural Information Processing Systems (NeurIPS)*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

Stąpor, K. (2018). Evaluating and comparing classifiers: Review, some recommendations and limitations. In M. Kurzynski, M. Wozniak, & R. Burduk (Eds.), *Proceedings of the 10th international conference on computer recognition systems cores 2017* (pp. 12–21). Springer International Publishing.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. *2017 IEEE International Conference on Computer Vision (ICCV)*, 843–852. https://doi.org/10.1109/ICCV.2017.97

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 3319–3328.

Sundararajan, M., Xu, J., Taly, A., Sayres, R., & Najmi, A. (2019). Exploring principled visualizations for deep network attributions. *IUI Workshops*.

Tahmassebi, A., Gandomi, A. H., McCann, I., Schulte, M. H. J., Goudriaan, A. E., & Meyer-Baese, A. (2018). Deep learning in medical imaging: Fmri big data analysis via convolutional neural networks. *Proceedings of the Practice and Experience on Advanced Research Computing*. https://doi.org/10.1145/3219104.3229250

van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., & Viergever, M. A. (2022). Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, *79*, 102470. https://doi.org/https://doi.org/10.1016/j.media.2022.102470

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., & Belongie, S. (2018). The inaturalist species classification and detection dataset. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8769–8778. https://doi.org/10.1109/CVPR.2018.00914

Waisberg, E., Ong, J., Kamran, S. A., Paladugu, P., Zaman, N., Lee, A. G., & Tavakkoli, A. (2023). Transfer learning as an AI-based solution to address limited datasets in space medicine. *Life Sciences in Space Research*, *36*, 36–38. https://doi.org/10.1016/j.lssr.2022.12.002

Xie, B., Lei, T., Wang, N., Cai, H., Xian, J., He, M., Zhang, L., & Xie, H. (2020). Computer-aided diagnosis for fetal brain ultrasound images using deep convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery*, *15*(8), 1303–1312. https://doi.org/10.1007/s11548-020-02182-3

Yan, S., Wang, C., Chen, W., & Lyu, J. (2022). Swin transformer-based gan for multimodal medical image translation. *Frontiers in Oncology*, *12*. https://doi.org/10.3389/fonc.2022.942511

Yang, Q., Zhu, X., Fwu, J.-K., Ye, Y., You, G., & Zhu, Y. (2021). Mfpp: Morphological fragmental perturbation pyramid for black-box model explanations. *2020 25th International Conference on Pattern Recognition (ICPR)*, 1376–1383. https://doi.org/10.1109/ICPR48806.2021.9413046

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision – eccv 2014* (pp. 818–833). Springer International Publishing.

Zhang, X., Zhang, Y., Zhang, G., Qiu, X., Tan, W., Yin, X., & Liao, L. (2022). Deep learning with radiomics for disease diagnosis and treatment: Challenges and potential. *Frontiers in Oncology*, *12*, 773840. https://doi.org/10.3389/fonc.2022.773840

Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. *International Conference on Learning Representations*. https://openreview.net/forum?id=BJ5UeU9xx