**HC2**

FYP Final Report

**Artificial Intelligence**

**Against**

**COVID-19**

by

CHEN, Siyu, ZHOU Taichang

**HC2**

Supervised by

Prof. CHEN Hao (陳浩)

Advised by

Luo Luyang and Zhuang Jiaxin

Submitted in partial fulfillment of the requirements for COMP4981 and CPEG4901

in the

Department of Computer Science and Department of Computer Engineering

The Hong Kong University of Science and Technology

2022-2023

Date of submission: April 20, 2023

**Abstract**

The COVID-19 pandemic has been growing worldwide and has caused millions of infections and deaths. When an outbreak occurs in a city, the exponential growth of confirmed cases would rapidly occupies the majority of medical resources, and exceptionally experienced doctors for diagnosing and treating patients with severe symptoms suspected of COVID-19. Although the testing technologies, including Polymerase Chain Reaction (PCR) Tests and Rapid Antigen Tests, have become much more mature, they are only indicators of the likelihood of a patient being infected with COVID-19. The infected patients still need to do further physical examinations like chest X-rays and CT scans to diagnose the severity of symptoms for future treatment.

In relief of the burden on the public medical system, many deep learning models are developed to support COVID-19 diagnosis with medical images such as chest X-rays or CT scans. Trained with the COVID-19 image datasets, the deep learning models can diagnose infection and even do semantic segmentation of lesions. The suggestions given by deep learning models can accelerate the decision-making process by doctors and reduce the possibility of mistakes and misdiagnoses caused by doctors due to the high workload during the outbreak.

To explore and exploit more potentials of deep learning models in COVID-19, we plan to explore more applications of Artificial Intelligence (AI) to medical images.

In this paper, **explainable artificial intelligence(XAI) methods** will be presented and applied to explain each neuron in the medical image processing network models for a better understanding of the network, and the model’s validity and accuracy will be evaluated as well.

We believe this direction is of great potential and worth exploring. Detailed explanations are presented as follows.

**Keywords: COVID-19, XAI, MRI, Medical Image, Post-hoc, Ante-hoc**

**Table of Contents**

[1 Introduction 7](#_Toc132917089)

[1.1 Explainable Artificial Intelligence (XAI) 7](#_Toc132917090)

[1.1.1 Background 7](#_Toc132917091)

[1.1.2 Interpretability 8](#_Toc132917092)

[1.1.3 Method 9](#_Toc132917093)

[1.2 Literature Survey 10](#_Toc132917094)

[1.2.1 Network Dissection[2], [3] 10](#_Toc132917095)

[1.2.2 Quantitative Testing with Concept Activation Vectors (TCAV)[4] 11](#_Toc132917096)

[2 Preparation 12](#_Toc132917097)

[2.1 Dataset 12](#_Toc132917098)

[2.1.1 Duke Breast Cancer MRI Dataset (MRI)[7] 12](#_Toc132917099)

[2.1.2 Diverse Dermatology Images Dataset (DDI)[8] 13](#_Toc132917100)

[2.1.3 Fitzpatrick17k Dataset[9] 14](#_Toc132917101)

[2.2 Model Fitting 15](#_Toc132917102)

[2.2.1 Model 15](#_Toc132917103)

[2.2.2 Fitting 16](#_Toc132917104)

[3 Methodology 18](#_Toc132917105)

[3.1.1 Post-hoc 18](#_Toc132917106)

[3.1.2 Ante-hoc 21](#_Toc132917107)

[4 Conclusion 23](#_Toc132917108)

[5 Project Planning 24](#_Toc132917109)

[5.1 Distribution of Work 24](#_Toc132917110)

[5.2 GANTT Chart 25](#_Toc132917111)

[6 Required Hardware &Software 26](#_Toc132917112)

[6.1 Hardware 26](#_Toc132917113)

[6.2 Software 26](#_Toc132917114)

[7 Appendix A: Meeting Minutes 27](#_Toc132917115)

[7.1 Minutes of the 1st Project Meeting 27](#_Toc132917116)

[7.2 Minutes of the 2nd Project Meeting 27](#_Toc132917117)

[7.3 Minutes of the 3rd Project Meeting 28](#_Toc132917118)

[7.4 Minutes of the 4th Project Meeting 28](#_Toc132917119)

[7.5 Minutes of the 5th Project Meeting 29](#_Toc132917120)

[7.6 Minutes of the 6th Project Meeting 29](#_Toc132917121)

[7.7 Minutes of the 6th Project Meeting 30](#_Toc132917122)

[8 References 31](#_Toc132917123)

# Introduction

## Explainable Artificial Intelligence (XAI)

**Explainable AI (XAI)**, also known as Interpretable AI, or Explainable Machine Learning (XML), is artificial intelligence (AI) in which humans can understand the reasoning behind decisions or predictions made by the AI.[1]

### Background

Complex ML models, such as Deep Neural Networks(DNN) models, are sometimes referred to as “**black box**” models because their mechanisms of making decisions are not explicitly accessible to human cognition. As a result, knowing how a model thinks or behave is desirable for the following reasons:

1. In terms of application for clinics, high transparency could easily build trust between patients and the models. Trust is directly related to the models’ professionalism or the level of understanding of the inner working of the model, not just a plain output such as classification labels or segmentation masks.
2. Knowing how a model thinks or behave allows doctors to understand when models are less reliable and to judge whether models are trustworthy based on facts.
3. Knowing how a model thinks or behave helps to identify critical perspectives of the data (e.g., the underlying biological mechanism) and could help the progress of research in both better understanding the subject matter and improving the model.

### Interpretability

In data mining and machine learning contexts, **interpretability** is defined as the ability to explain or present intelligible terms to humans. The interpretability of machine learning models can be generally divided into 2 categories:

* **Ante-hoc Interpretability**

It refers to making the model itself have interpretability by training a model with a simple structure and good interpretability or a self-explanatory model that combines interpretability into a specific model structure.

* **Post-hoc Interpretability**

It refers to explaining trained machine learning models by developing interpretability techniques. According to different interpretation goals and objects, post-hoc interpretability can be divided into global and local interpretability. Global interpretability aims to help people understand the overall logic behind the complex model and the internal working mechanism, and local interpretability aims to help people understand the decision-making process and basis for each input sample of the machine learning model.

图示

描述已自动生成

Figure : Workflow of Ante-hoc and Post-hoc interpretability.

### Method

So basically, we can roughly understand that the AI model with interpretability is XAI. According to the two types of interpretability, to achieve the goal that we want to design a medical image XAI, we can just try to show our AI model has one of these two interpretability. In another word, we have two main ways to build an XAI.

In this work, we follow the same workflow in Figure 1 and show how we combined and improved some of the current methodologies to propose our methodology. Our project explore and tried both Post-hoc and Ante-hoc ways.

## Literature Survey

Before we introduce our project design and methodology, let’s show some related works in terms of Post-hoc and Ante-hoc interpretability. The following papers and designs are the inspiration for our projects.

### Network Dissection[2], [3]

Network dissection is a very intuitive Post-hoc methodology to explain how an ML model thinks by explaining what concepts each layer of the model might be associated with. Since we believe that while training a model, the model could learn something important and store the memory into the weight and bias of its layers. So network dissection tries to extract this knowledge and match them with a bunch of common concepts that humans can understand. As shown in the D part in Figure 2, layer conv5\_3 can be related to such amount of concepts.

图表

中度可信度描述已自动生成

Figure : Sample outputs of Network dissection on aVGG-16 model.

### Quantitative Testing with Concept Activation Vectors (TCAV)[4]

TCAV is another intuitive Post-hoc methodology to explain how an image classification network thinks by explaining what concepts each unit in one layer of the model might be related to. This methodology and its improved version[5] mainly treat each concept as a separate multi-label classification problem so that they can use different regression methods to discover the positive/negative association of each unit with each label of each concept.

图示

描述已自动生成

Figure : A simple binary classification problem.

# Preparation

## Dataset

After serious consideration of the problem that we lack a COVID-19 dataset with reliable clinical concept marks for each image, we decided to use other similar medical images, including both 2D and 3D images, to validate our methodologies.

For each dataset, we did the data augmentation method[6], which is a technique of artificially increasing the training set by creating modified copies of a dataset using existing data. It includes image change methods such as different axis and angle rotation, horizontal flip, padding, crop, and translation.

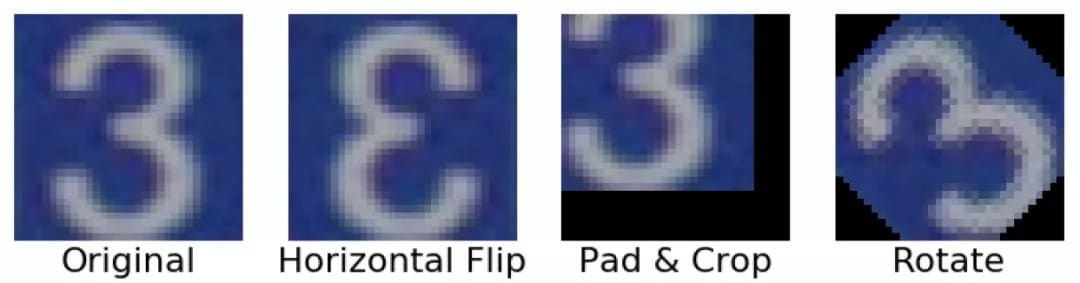


Figure : A sample of the data augmentation method.

### Duke Breast Cancer MRI Dataset (MRI)[7]

The Duke Breast Cancer MRI Dataset is a 3D MRI dataset, where lesions will hold a noticeable brightness increment compared to common tissues. This is similar to the images for the performance of COVID-19 on the lungs. The dataset contains about 922 scanning results from 2000 to 2014, with over 50 clinical-related concepts like Nottingham Grade. Since the results hold different scanning numbers as they are coming from different scanning machines, we only choose three scans as three channels from each MRI, which are original images before contrast media injection, first scan images with low magnetic field intensity after contrast media injection, second scan images with higher intensity. After discarding several unreadable images, patching lesions from images, and dividing the datasets into training sets and testing sets at a ratio of about 2:1 randomly, we finalized a dataset containing 908 MRI images in the shape of (3, 100, 100, 100).

图片包含 图示

描述已自动生成

Figure : The column explanations on the Duke Breast Cancer MRI dataset. Each column is a clinical concept regarding breast cancer.

### Diverse Dermatology Images Dataset (DDI)[8]

The Diverse Dermatology Images is a biopsy-proven skin disease dataset with diverse skin tone representation, where all images resemble natural images that are taken by cameras. The dataset contains 656 images representing 570 unique patients with 48 professional clinical labels, and it is divided into a training set and a testing set in the ratio of 8:2.

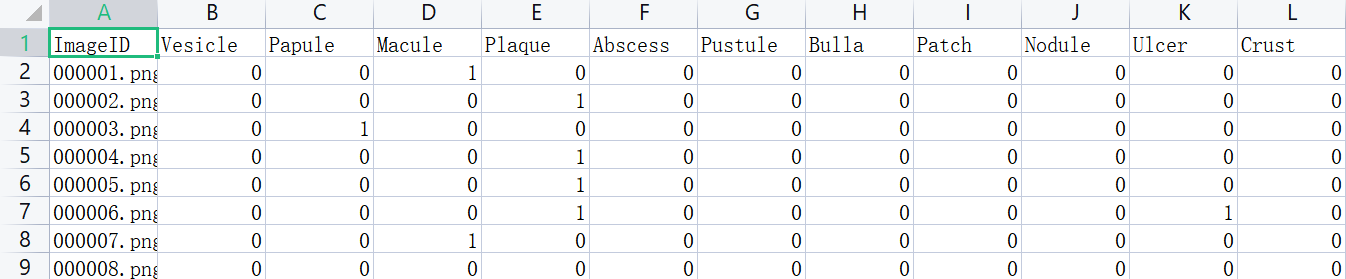


Figure : The clinical features of the Diverse Dermatology Image dataset. Each column is a clinical concept regarding skin disease.

### Fitzpatrick17k Dataset[9]

The Fitzpatrick17k dataset is a skin disease dataset similar to Diverse Dermatology Images but with more skin conditions: 114 specific conditions, nine middle-scale classes, three comprehensive classes, and a much larger scale: this dataset contains 16,577 clinical images. Same as the Diverse Dermatology Images, we divided this dataset into train and test in the ratio of 8:2. However, since the original images lack clinical labels, the labels from Skincon dataset for these images only involve 3690 images, and the dataset smaller when it comes to our concept corrected model.

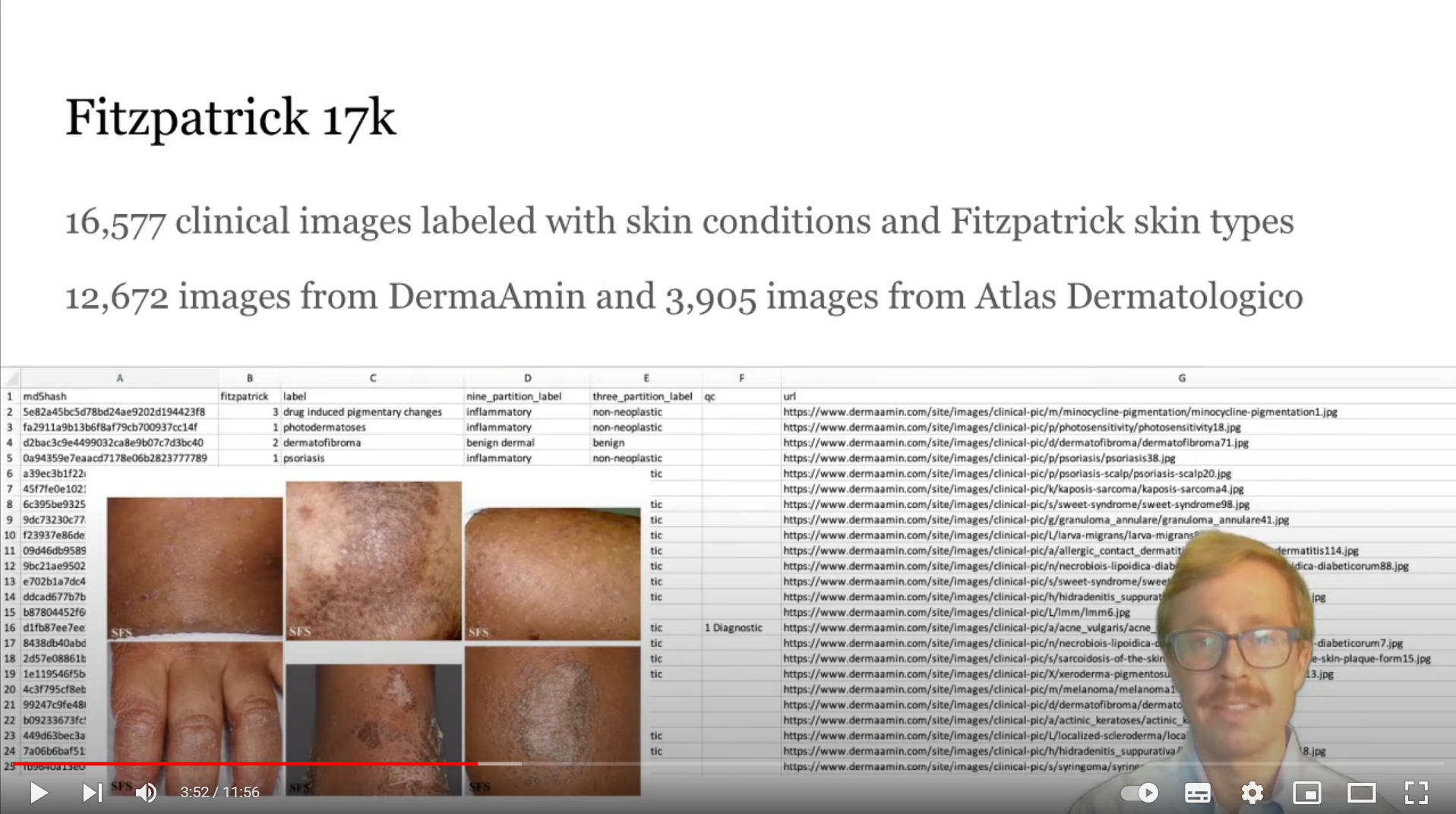


Figure : Sample images and related skin disease classes from Fitzpatrick17k Dataset.

## Model Fitting

Since we mainly use the Duke Breast Cancer MRI dataset to implement model fitting while the other two datasets are used as additional methodology validation datasets, we will mainly talk about the construction and performance of the model based on the Duke Breast Cancer MRI dataset in this part.

### Model

The main task of this model has an academic concept called molecular typing of cancer. Molecular subtyping of cancer refers to the use of omics data to find clusters of tumors within a cancer type that have shared characteristics. In other words, this model tries to classify breast cancer into different groups based on molecular data and classification categories.

We mainly use two model structures: ResNet and DenseNet.

Residual Neural Network(ResNet) is a specific type of neural network that was introduced in 2015. The residual connection between layers in this network provides another path for data to reach the latter layers of the neural network by skipping some layers, as shown in Figure 8.[10]

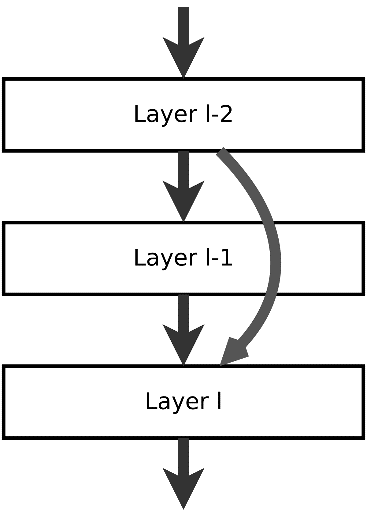


Figure : A simple ResNet structure.

Dense Convolutional Network(DenseNet) is another type of neural network that has a similar idea to ResNet. The difference is that the former layers in DenseNet will have a dense connection with each layer afterward. In other words, all layers will be directly connected, as shown in Figure 9.[11]

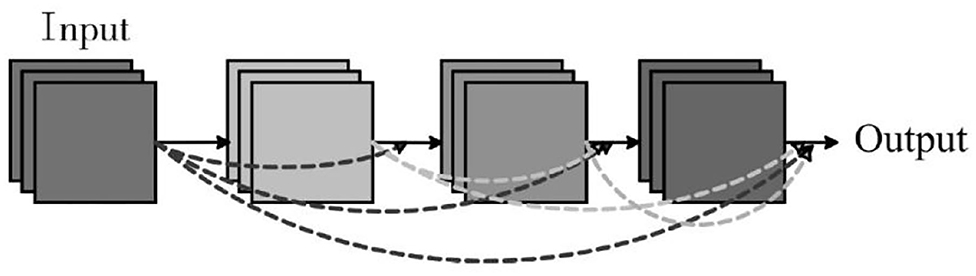


Figure : A simple DenseNet structure.

Depending on the number of convolutional layers, we can add different suffixes to different neural networks. For example, ResNet18, ResNet34, and DenseNet121.

### Fitting

In this project, we used the following neural network structure:

* ResNet18, ResNet34, ResNet50, ResNet101, ResNet152.
* DenseNet121, DenseNet161, DenseNet169, DenseNet201.

After training those models with the Duke Breast Cancer MRI dataset, we found that the DenseNet121 has the best performance in terms of the measurement of Area under the Receiver operating characteristic(ROC) Curve(AUC).

图表, 折线图, 散点图

描述已自动生成

Figure : Model training history of DenseNet121 for MRI on class PR

Our model could reach AUC up to about 0.63 on the binary class progesterone receptors(PR), while the dataset-provider’s model on PR is about 0.62.

# Methodology

Inspired by the literature survey we learned, we propose our Post-hoc and Ante-hoc methodology. The Post-hoc explanation is extracted after the model training, which aims at disclosing the black-box model. On the other side, the Ante-hoc explanation is to convert those originally covered parts to explainable parts, like changing the linear parts to decision trees, or ensuring the model learns certain concepts.

### Post-hoc

We have used the method called TCAV[4] to disclose the concepts learned by the model’s convolutional part, using SVM to regress the feature map to clinical concepts.

1. To calculate the Concept Activation Vectors (CAV), we first calculate the feature map of the last layer of the convolutional part. The feature map is the intermediate output result from a certain convolutional layer given an input picture. To get the feature map, we have used the hook() function provided by Pytorch or saved the output as an attribute of the model class.

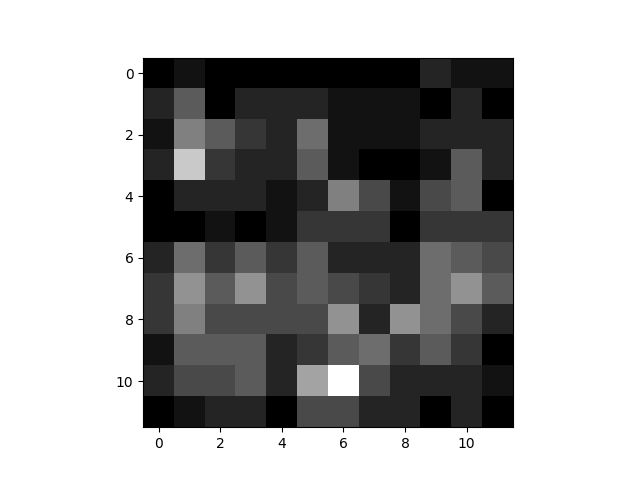


Figure : A sample piece of the feature map.

1. We feed a sample from our feature maps including an equal number of positive or negative for the concept and the corresponding clinical concept to a regression model, including SVM, Linear Regression, and Logistic Regression. Results from different models hold noticeable consistency, with a little ranking switch between 2 concepts. The weight for each corresponding pixel is viewed as the activation of the pixel to the concept, as the higher the weight is, the more sensitive to this pixel. To filter those noise and error activations, we set a threshold of the top 1% for all concept activations.

图表, 树状图

描述已自动生成 图片包含 日历

描述已自动生成

Figure : Sample of pieces of the activation map, different colors represent different clinical concepts.

1. We choose the most activated concept as the corresponding concept for the pixel, and we count the number of activating pixels for each concept to find the most important concepts. The concept here contains many medical concepts, which are used for diagnosis by doctors, while some irrelevant concepts are also included. We believe our result is acceptable due to the limitation of the model performance and task difficulty. All medical data retrieved will be used in the Ante-hoc part.

|  |  |
| --- | --- |
| MRI (ER) | DDI |
| |  |  | | --- | --- | | **Concepts** | **Weight** | | Multicentric/Multifocal | 1817 | | Known Ovarian Status | 1699 | | Field Strength (Tesla) | 1604 | | Lymphadenopathy or Suspicious Nodes | 1464 | | Mol Subtype | 970 | | Nottingham grade | 960 | | Contralateral Breast Involvement | 747 | | Patient Position During MRI | 698 | | Slice Thickness | 554 | | Manufacturer | 467 | | |  |  | | --- | --- | | **Concepts** | **Weight** | | Brown (Hyperpigmentation) | 10507 | | Erythema | 9158 | | Papule | 4527 | | Scale | 4162 | | Plaque | 3029 | | Dome-shaped | 1909 | | White (Hypopigenmentation) | 563 | | Black | 280 | | Crust | 227 | | Nodule | 112 | |

Figure : Top 10 Important Concepts for different datasets.

### Ante-hoc

After obtaining the important concepts, we manually pick some clinical concepts as there are some irrelevant concepts included like the manufacturer of the MRI machine. Then we put the concepts to a Concept Correction, a logistic regression to generate those selected concepts, to ensure that the convolutional layers learn selected concepts. Additionally, to make sure that the model learns different concepts, we add a similarity penalty to the feature map. The loss function that we are trying to optimize is:

Equation (1):

Equation (2):

Equation (3):

Where:

: ith concept.

: input images.

: ground truth.

: model prediction.

: Cross Entropy Loss, Logistic: Logistic regression.

, : hyperparameter controls the weight of loss concepts/similarities.

: ith vector extracted from feature map where pixels are at the same position but in different channels.

# Conclusion

In this project, we have trained deep learning models on different medical datasets with clinical concepts and have successfully presented an explanation system including two explanation methods established from TCAV and Logistic Regression, to reveal the black box of deep learning models and to build more trustable transparent deep learning models.

According to our experiment results, we believe that our Post-hoc method is able to extract clinical concepts that are significant for prediction tasks, and our Ante-hoc method can ensure the model make trustable decisions based on reliable medical concepts, just like doctors diagnose based on expert knowledge. Additionally, with transparency improved, the model performance still retains, or even be improved a little bit (1%).

There are some experiments and methods that can be further explored or improved. For example, our method of determining important concepts is direct and simple, there may be some more ingenious methods to gain those concepts more accurately. Additionally, the method used to involve selected concepts can also be improved so that the model can be more reliable with much better performances.

We believe that our work on this Final Year Project is not only to demonstrate our understanding in the area of deep learning but also to provide us with a chance to help those in need with our effort and knowledge. Since 2019, COVID-19 reveals the potential risk and medical resource shortages all around the world, we help our work could reduce the workload of doctors, assist with more precise diagnoses, and provide advice to those who cannot achieve enough health services.

# Project Planning

## Distribution of Work

|  |  |  |
| --- | --- | --- |
| **Task** | **ZHOU Taichang** | **CHEN Siyu** |
| Literature Survey | ○ | ● |
| Data Preparation | ● | ○ |
| Data Augmentation of Dataset | ● | ○ |
| Model Fitting | ○ | ● |
| Feature Map Extraction(Post-hoc) | ● | ○ |
| Activation Map Extraction(Post-hoc) | ● | ○ |
| Pixel Concepts Extraction(Post-hoc) | ○ | ● |
| Concepts Correction(Ante-hoc) | ● | ○ |
| Proposal/Report Writing | ○ | ● |

● Leader ○ Assistant

## GANTT Chart

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Task | July | Aug | Sep | Oct | Nov | Dec | Jan | Feb | Mar | Apr |
| Do the literature survey |  |  |  |  |  |  |  |  |  |  |
| Analyze related works |  |  |  |  |  |  |  |  |  |  |
| Dataset Preparation |  |  |  |  |  |  |  |  |  |  |
| Data Augmentation of Dataset |  |  |  |  |  |  |  |  |  |  |
| Model Construction |  |  |  |  |  |  |  |  |  |  |
| Model Fitting |  |  |  |  |  |  |  |  |  |  |
| Model Visualization |  |  |  |  |  |  |  |  |  |  |
| Model Evaluation |  |  |  |  |  |  |  |  |  |  |
| Feature Map Extraction(Post-hoc) |  |  |  |  |  |  |  |  |  |  |
| Feature Map Visualization |  |  |  |  |  |  |  |  |  |  |
| Activation Map Extraction(Post-hoc) |  |  |  |  |  |  |  |  |  |  |
| Activation Map Visualization |  |  |  |  |  |  |  |  |  |  |
| Pixel Concepts Extraction(Post-hoc) |  |  |  |  |  |  |  |  |  |  |
| Pixel Concepts Visualizations |  |  |  |  |  |  |  |  |  |  |
| Concepts Correction(Ante-hoc) |  |  |  |  |  |  |  |  |  |  |
| Write the proposal |  |  |  |  |  |  |  |  |  |  |
| Write the monthly reports |  |  |  |  |  |  |  |  |  |  |
| Write the progress report |  |  |  |  |  |  |  |  |  |  |
| Write the final report |  |  |  |  |  |  |  |  |  |  |
| Prepare for the presentation |  |  |  |  |  |  |  |  |  |  |
| Design the project poster |  |  |  |  |  |  |  |  |  |  |

# Required Hardware &Software

## Hardware

Development PC: PC with MS Windows 10

Remote Server: Larger CPU/GPU for training model

## Software

MobaXterm: For connection with Remote Server

Visual Studio Code For code visualization

Python Programming languages

Anaconda, miniconda Package management system and deployment

# Appendix A: Meeting Minutes

## Minutes of the 1st Project Meeting

**Date**: Sep 4, 2022

**Time**: 3:00 pm

**Place**: ZOOM

**Present**:  CHEN Siyu; ZHOU Taichang; Prof. CHEN Hao; Luo Luyang; and Zhuang Jiaxin.

1. Talk about the background, knowledge learned, and research interests.
2. Report on progress and current ideas of medical image analysis of COVID-19.
3. Determine the research directions: XAI.
4. Divide into sub-groups and assign advisers to sub-groups.
5. Make timelines and deadlines for each step of the FYP.

## Minutes of the 2nd Project Meeting

**Date**: Sep 13, 2022

**Time**: 9:30 am

**Place**: ZOOM

**Present**:  CHEN,Siyu, ZHOU Taichang, Luo Luyang.

1. Report on XAI direction progress.
2. Determine the detailed research direction.
3. Determine a 4-step plan for our project.
4. Talk about how to write the FYP proposal.

## Minutes of the 3rd Project Meeting

**Date**: Sep 29, 2022

**Time**: 8:00 pm

**Place**: ZOOM

**Present**:  CHEN Siyu; ZHOU Taichang; Prof. CHEN Hao; Luo Luyang; and Zhuang Jiaxin.

1. Report on XAI direction progress.
2. Talked about the literature survey.
3. Ask for more help on the dataset finding other than the Duke dataset.

## Minutes of the 4th Project Meeting

**Date**: Nov 3, 2022

**Time**: 6:00 pm

**Place**: ZOOM

**Present**:  CHEN Siyu; ZHOU Taichang; Prof. CHEN Hao; Luo Luyang; and Zhuang Jiaxin.

1. Report on XAI direction progress.
2. Evaluation of our current Post-hoc methodology.
3. Method network dissection itself cannot solve the problem so we may need to combine other methodologies such as causal analysis.
4. Feedback about the monthly report.

## Minutes of the 5th Project Meeting

**Date**: Jan 13, 2023

**Time**: 8:00 pm

**Place**: ZOOM

**Present**:  CHEN Siyu; ZHOU Taichang; Prof. CHEN Hao; Luo Luyang; and Zhuang Jiaxin.

1. Report on XAI direction progress.
2. Talked about our winter semester plans and outcomes.
3. We need to speed up to get more outcomes.
4. Conclude that Post-hoc direction do can find something new and important.

## Minutes of the 6th Project Meeting

**Date**: Feb 11, 2023

**Time**: 6:00 pm

**Place**: ZOOM

**Present**:  CHEN Siyu; ZHOU Taichang; Prof. CHEN Hao; Luo Luyang; and Zhuang Jiaxin.

1. Report on XAI direction progress.
2. Post-hoc direction is about to finish. Outcome Visualization.
3. Literature Survey on Ante-hoc direction.
4. We need to apply our method to more datasets to evaluate whether our methodology is meaningful and repeatable.

## Minutes of the 6th Project Meeting

**Date**: Apr 12, 2023

**Time**: 8:00 pm

**Place**: ZOOM

**Present**:  CHEN Siyu; ZHOU Taichang; Prof. CHEN Hao; Luo Luyang; and Zhuang Jiaxin.

1. Report on XAI direction’s outcome and contribution.
2. Get Some advice from supervisors and advisors.
3. Determine whether to continue to work on it after FYP.
4. FYP final report and plan.
5. Presentation rehersal.

# References

[1] P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, and M. A. Przybocki, “Four Principles of Explainable Artificial Intelligence,” 2020, doi: 10.6028/nist.ir.8312-draft.

[2] B. Zhou, D. Bau, A. Oliva, and A. Torralba, “Interpreting Deep Visual Representations via Network Dissection,” *Ieee T Pattern Anal*, vol. 41, no. 9, pp. 2131–2145, 2017, doi: 10.1109/tpami.2018.2858759.

[3] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, and A. Torralba, “Understanding the role of individual units in a deep neural network,” *Proc National Acad Sci*, vol. 117, no. 48, pp. 30071–30078, 2020, doi: 10.1073/pnas.1907375117.

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

[5] S. Singla, S. Wallace, S. Triantafillou, and K. Batmanghelich, “Using Causal Analysis for Conceptual Deep Learning Explanation,” *Arxiv*, 2021, doi: 10.48550/arxiv.2107.06098.

[6] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *J Big Data*, vol. 6, no. 1, p. 60, 2019, doi: 10.1186/s40537-019-0197-0.

[7] A. Saha *et al.*, “A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features,” *Brit J Cancer*, vol. 119, no. 4, pp. 508–516, 2018, doi: 10.1038/s41416-018-0185-8.

[8] R. Daneshjou *et al.*, “Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set,” *Arxiv*, 2022, doi: 10.48550/arxiv.2203.08807.

[9] M. Groh *et al.*, “Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset,” *Arxiv*, 2021, doi: 10.48550/arxiv.2104.09957.

[10] S. Targ, D. Almeida, and K. Lyman, “Resnet in Resnet: Generalizing Residual Architectures,” *Arxiv*, 2016, doi: 10.48550/arxiv.1603.08029.

[11] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, “DenseNet: Implementing Efficient ConvNet Descriptor Pyramids,” *Arxiv*, 2014, doi: 10.48550/arxiv.1404.1869.