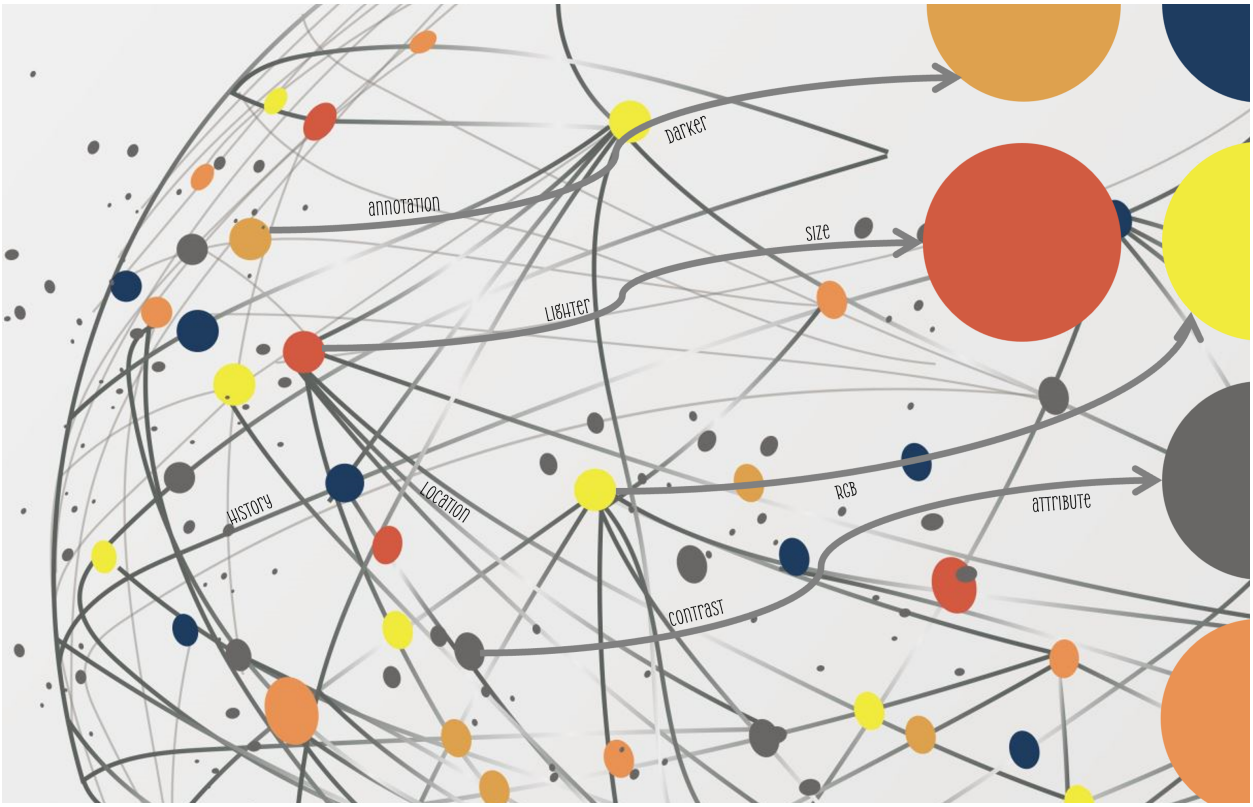
CHALMERS
UNIVERSITY OF TECHNOLOGY



Robust Medical Image Analysis using Privileged Information

Master's Thesis

Master's thesis in Biomedical Engineering

APALA CHAKRABARTI

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2024
www.chalmers.se

Master's thesis 2024

Robust Medical Image Analysis using Privileged Information

A Study on the Interplay of Privileged Information of Domain Adaptation in the Context of Medical Imaging

APALA CHAKRABARTI

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Division of Signal processing and Biomedical Engineering
Chalmers University of Technology
Gothenburg, Sweden 2024

Robust Medical Image Analysis using Privileged Information
A Study on the Interplay of Privileged Information of Domain Adaptation in the Context of Medical Imaging
APALA CHAKRABARTI

Supervisor: Fredrik Johansson, Department of Computer Science & Engineering
Examiner: Ida Häggström, Department of Electrical Engineering

Master's Thesis 2024
Department of Electrical Engineering
Division of Signal processing and Biomedical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2024

Robust Medical Image Analysis using Privileged Information
A Study on the Interplay of Privileged Information of Domain Adaptation in the Context of Medical Imaging
APALA CHAKRABARTI Department of Electrical Engineering
Chalmers University of Technology

Abstract

Domain adaptation is a crucial task in medical diagnosis and treatment planning, as it enables models trained on large, labeled datasets to be effectively applied to smaller, domain-specific datasets. This is particularly challenging due to data scarcity and shifts in data distribution. Privileged Information (PI), such as binary attributes or bounding boxes, has the potential to improve machine learning models' adaptability across diverse domains. This study aims to investigate the role of PI in domain adaptation for medical image classification. The results of this experiment indicate that integrating PI led to increased accuracy and stabilized prediction accuracies. Furthermore, the findings affirm the importance of both the quantity and correlation of the PI provided and its correlation with output labels in enhancing model performance, thereby supporting the fundamental principles of domain adaptation. Moreover, the study underscores the significance of strategically considering PI attributes during model training to achieve stable output accuracy and effectively mitigate domain shift. This comprehensive study will help improve diagnostic accuracy in various domains, especially healthcare, which can lead to more effective treatments and better patient outcomes.

Keywords: Machine learning, Domain adaptation, Medical image classification, Privileged Information.

Acknowledgements

This thesis would not have been possible without the support and encouragement of many people. I would like to express my deepest gratitude to my supervisor, Dr. Fredrick Johansson, for his invaluable guidance, patience, and insightful feedback throughout this research journey. His expertise and dedication have been instrumental in the completion of this work.

I am also thankful to my examiner, Dr. Ida Häggström, for her constructive suggestions and support. Her comments have greatly improved the quality of this thesis. Special thanks to Alba for providing valuable opposition and feedback that have enriched this work.

I extend my heartfelt thanks to my friends José, Kelly, Sara, Eduardo, Leandros, Emil, Albin, and Rodrigo for their unwavering support over the last year. Your constructive feedback, comments, and encouragement have been crucial in shaping this thesis. Your friendship has been a source of strength and motivation.

Moreover, I am deeply grateful to Manavi, Devarchita and Vardan for their support from all over. Your belief in me, motivation, and constant encouragement have been truly inspiring. Without your support, this thesis would not have been possible. Thank you for always having my back.

Finally, my heartfelt thanks go to my parents, Baba and Ma. Without your infinite patience, love, and strength, completing this thesis would not have been possible. Your support has been my foundation, and I am forever grateful for everything you have done.

Thank you for your contributions to this thesis. This achievement would not have been possible without your support.

Apala Chakrabarti, Gothenburg, June, 2024

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

ACML	Attributes Ranked by Correlation with Male Label
AOC	Attributes Ranked by Overall Correlation
AT	Atelectasis
AUC	Area Under the Curve
CAGR	Compound Annual Growth Rate
CelebA	CelebFaces Attributes Dataset
CM	Cardiomegaly
CNN	Convolutional Neural Network
CT	Computed Tomography
DALUPI	Unsupervised Domain Adaptation by Learning using Privileged Information
DANN	Domain-Adversarial Neural Network
fMRI	Functional Magnetic Resonance Imaging
FPR	False Positive Rate
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NIH	National Institutes of Health Clinical Centre
NLP	Natural Language Processing
PCA	Principal Component Analysis
PE	Pleural Effusion
PET	Positron Emission Tomography
PI	Privileged Information
R-CNN	Region-based Convolutional Neural Network
ResNet	Residual Network
RGB	Red-Green-Blue
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SDA	Supervised Domain Adaptation
S-SDA	Semi-supervised domain adaptation
SML	Supervised Machine Learning
SPECT	Single Photon Emission Computed Tomography
SSML	Semi-Supervised Machine Learning
SVM	Support Vector Machines
TPR	True Positive Rate
UDA	Unsupervised Domain Adaptation
UML	Unsupervised Machine Learning

Nomenclature

Below is the nomenclature of indices, sets, parameters, and variables that have been used throughout this thesis.

Indices

i	Index for sample in dataset
t	Index for iteration

Sets

$\mathcal{N}$	Set of samples in dataset

Parameters

θ	Parameters of the model
N	Size of the dataset
y_i	Actual label for the ith sample
α	Learning rate, which controls the step size of the weight updates
λ	Trade-off parameter controlling the importance of domain adaptation objective

Variables

$\hat{y}_i$	Predicted output for the ith sample
$\mathcal{L}(\theta)$	Loss function over the dataset
$L(\hat{y}_i, y_i)$	Discrepancy between predicted and actual values for a single sample
w_t	Current weight at iteration t

w_{t+1}	Updated weight at iteration $t+1$
$\frac{\partial \mathcal{L}}{\partial w}$	Gradient of the loss function with respect to the weights
x	Input to the residual block and output from the previous layer
$H(x)$	Output of the residual block
$F(x)$	Function representing the residual mapping learned by the network
G	Feature extractor
D_{domain}	Domain classifier
D_{class}	Discriminative (task-specific) classifier
h_f	Extracted features from the input data
$\mathcal{L}_{\text{task}}$	Loss function for the task-specific classification
$\mathcal{L}_{\text{domain}}$	Loss function for domain classification
$P(\text{class}\|x)$	Probability of class label given input x
$P(\text{source}\|x)$	Probability of source domain given input x

Contents

List of Acronyms ix

Nomenclature xii

List of Figures xix

List of Tables xxi

1 Introduction 1

2 Background 5
2.1 Medical imaging . . . 5
2.1.1 Medical Imaging Modalities . . . 5
2.2 Machine Learning in Medical Imaging . . . 6
2.2.1 Input, Output and Loss . . . 8
2.2.2 Convolutional Neural Network (CNN) . . . 9
2.2.2.1 Residual Network (ResNet) . . . 11
2.2.2.2 Region-based Convolutional Neural Network (R-CNN) 12
2.2.2.3 Fast Region-based Convolutional Neural Network (Fast R-CNN) . . . 13
2.2.3 Challenges Faced in Imaging . . . 13
2.3 Domain Adaptation Challenges . . . 14
2.3.1 Domain-Adversarial Neural Network . . . 14
2.4 Privileged Information . . . 16

3 Methods 19
3.1 Softwares and Tools . . . 19
3.2 DALUPI Method . . . 19
3.2.1 Two-stage approach . . . 21
3.3 Experiment Design . . . 22
3.4 Experiment I: Binary Attributes . . . 23
3.4.1 Hypotheses: Attributes . . . 24
3.4.2 CelebA Dataset . . . 24
3.4.3 Attribute-Label Correlation . . . 24
3.4.4 Image Preprocessing . . . 26
3.4.5 Model Configurations . . . 27
3.5 Experiment II: Bounding boxes . . . 27

3.5.1 Hypotheses: Bounding box . . . 27
3.5.2 Dataset 1: NIH Chest X-Ray . . . 27
3.5.3 Dataset 2: CheXpert . . . 28
3.5.4 Bounding Box Estimation . . . 28
3.5.5 Data Preprocessing . . . 30
3.5.6 Model Adaptations . . . 30
3.6 Evaluation Metrics . . . 31
3.6.1 Accuracy Average . . . 31
3.6.2 Receiver Operating Characteristic curve . . . 31
3.6.3 Area Under the ROC Curve (AUC) . . . 31

4 Results **33**
4.1 Binary Attributes as PI . . . 33
4.1.1 Accuracy vs Number of Attributes: Visualisation . . . 33
4.1.1.1 Attributes in Order of the ACML List . . . 33
4.1.1.2 Attributes in Order of the AOC List . . . 35
4.1.2 DALUPI vs Baseline Algorithms . . . 37
4.1.2.1 SL-S, DANN and MDD Performance Analysis . . . 37
4.1.2.2 DALUPI Performance Analysis . . . 37
4.1.3 Observations: Experiment I . . . 38
4.2 Bounding Boxes as PI . . . 38
4.2.1 Estimation of Bounding Boxes . . . 38
4.2.2 AUC of Different Bounding Box Sizes . . . 39

5 Discussions **43**
5.1 Experiment-Specific Inferences . . . 43
5.2 Overall Findings and Implications . . . 44
5.3 Limitations and Future Scope . . . 45
5.3.1 Generalization to Other Domains . . . 45
5.3.2 Scalability . . . 45
5.3.3 Noise and Variability . . . 45
5.3.4 Scope of PI . . . 45
5.3.5 Assumption of Domain Shift . . . 45
5.4 Societal, Ethical and Ecological considerations . . . 46
5.4.1 Societal Impact . . . 46
5.4.2 Ethical Considerations . . . 46
5.4.3 Ecological Impact . . . 47

6 Conclusion **49**

Bibliography **51**

A Appendix 1 **I**
A.1 Types of ML . . . I
A.1.1 Supervised Machine Learning . . . I
A.1.1.1 Regression . . . I
A.1.1.2 Classification . . . I

A.1.2 Unsupervised Machine Learning II
A.1.2.1 Clustering II
A.1.3 Semi-Supervised Machine Learning II
A.1.3.1 Self-training II
A.1.3.2 Co-training II
A.2 Chest X-Ray labels III
A.2.1 Atelectasis III
A.2.2 Pleural Effusion IV
A.2.3 Cardiomegaly IV
A.3 Attributes and Correlations IV

List of Figures

2.1 Simplified block diagram of a ML model with input x and output y . 8
2.2 Generalised CNN Architecture 10
2.3 Generalised ResNet Architecture 11
2.4 Generalised R-CNN Architecture 12
2.5 Generalised Fast R-CNN Architecture 13
2.6 Generalised DANN Architecture 15

3.1 The presence and absense of domain adaptation between source and target domains 20
3.2 Block diagram depicting the two-stage approach of DALUPI 22
3.3 Experiment Design 23
3.4 Sample images from CelebA with corresponding attributes. Source: [39] 25
3.5 Sample images from NIH Chest X-ray showing common thorax diseases as taken directly from the data source. Source: [42] 28
3.6 Sample images from CheXpert of a patient with Pleural Effusion as taken directly from the data source. Source: [43] 29
3.7 Typical ROC curve with TPR vs FPR at different thresholds 32

4.1 Source accuracy plot for DALUPI according to ACML order of attributes 34
4.2 Target accuracy plot for DALUPI according to ACML order of attributes 35
4.3 Source accuracy plot for DALUPI according to AOC order of attributes 36
4.4 Target accuracy plot for DALUPI according to AOC order of attributes 36
4.5 Different sizes of bounding boxes (x, 2x and 4x) 39

A.1 Atelectasis, Right Lower Lobe. Source: [47] III

List of Tables

2.1 Overview of Different Layer Types in CNNs 9
2.4.1 Types of PI seen in Medical Imaging 17

3.3.1 Features, Datasets, and Descriptions 23
3.4.2 Attributes and Correlation to Label: Male 25
3.4.3 Attributes and Overall Correlation 26
3.4.4 Hyperparameters 26
3.5.5 Comparison of labels from CheXpert and NIH Chest X-ray 29

4.1.1 Comparison of average accuracy of DALUPI vs other models 37
4.2.2 Label Specific AUC over Different Bounding Box Sizes in the Source Domain 39
4.2.3 Label Specific AUC over different Bounding Box Sizes in the Target Domain 40
4.2.4 Comparison of average accuracy of DALUPI vs other models 41
4.2.5 Source accuracy when bounding box locations are moved 42
4.2.6 Target accuracy when bounding box locations are moved 42

A.3.1Attribute Correlation V

1
Introduction

Accurate diagnosis based on medical imaging plays a pivotal role in clinical decision-making, influencing treatment strategies and patient outcomes. However, despite significant advancements in medical imaging technology, diagnostic errors remain a challenge, with studies consistently showing missed findings rates and diagnostic errors by experienced radiologists [1]. These errors not only impact patient care but also contribute significantly to medical malpractice claims against radiologists.

Various methods have been developed to aid radiologists in diagnosis, including visualization techniques enhancement, image fusion, and telemedicine platforms. Despite these advancements, diagnostic outcomes have not seen significant improvements, prompting the need for more reliable and robust diagnostic tools. In response to this challenge, the integration of machine learning (ML) algorithms has emerged as a promising approach to augment radiologists' capabilities in interpreting medical images and improving diagnostic accuracy.

ML models, which can establish patterns or make decisions based on previously unseen data, have gained widespread traction across various industries, including healthcare. However, challenges persist in developing robust algorithms capable of accommodating images from diverse sources due to the absence of standardized guidelines for image resolution, angles, and other imaging parameters, along with slight variations in images from different instruments [2, 3].

One of the primary challenges in medical image analysis is domain shift, where images from different sources have varying distributions, leading to suboptimal performance of ML models trained on one dataset when applied to another. Domain adaptation refers to the process of modifying models to handle the differences between source and target domains effectively. To address this challenge, novel approaches are emerging that leverage PI – additional data accessible solely during training, enriching the learning process while not influencing predictions at test time [4]. These approaches aim to capitalize on available information to enhance model generalization and adaptability in the face of distributional changes. Magnetic Resonance (MR) images, known for their high contrast and detailed soft tissue imaging capabilities, present unique challenges and opportunities in domain adaptation due to variations in scanning protocols and hardware.

Prior studies in medical image analysis have investigated diverse domain adaptation techniques for enhanced model performance. Adversarial domain adaptation

[5] aims to mitigate variability in MR image segmentation across imaging protocols. Reviews [6] have surveyed domain adaptation methods for histopathology image analysis. Unsupervised domain adaptation [7] addresses data heterogeneity in brain lesion segmentation. Challenges include domain-specific sensitivities [5][7], anatomical dependency [5], and parameter tuning [6][7]. Similarly, the use of PI in the classification of medical images is highlighted by significant studies as seen in literature. In brain disorder diagnosis, combining structural MRI (sMRI) as unprivileged data with functional MRI (fMRI) as PI offers accuracy gains, though reconciling modalities poses challenges [8]. Alzheimer's disease diagnosis employs ensemble privileged learning with MRI data as unprivileged and undisclosed PI, highlighting potential but acknowledging integration hurdles [9]. Ultrasound elastography employs PI for strain reconstruction, enhancing accuracy but implying seamless data fusion [10]. Glioma grading employs multimodal convolutional neural networks with privileged learning for improved accuracy, facing challenges in effective privileged data integration [11]. These studies demonstrate the potential of PI while also recognizing the difficulties posed by modalities, mismatches, and data distribution. Generalizing domain adaptation and the use of PI to clinical scenarios demands further exploration and advancement.

Previous research highlights certain challenges in medical image analysis, encompassing domain shift from diverse imaging protocols, anatomical variations impacting segmentation, and data heterogeneity. Integrating multiple modalities faces difficulties arising from mismatches and limited labeled data. Similarly, leveraging PI effectively without overfitting presents challenges. To address this gap, a compelling approach involves integrating both domain adaptation and PI to enhance robustness, thereby facilitating practical solutions for medical applications [12]. In this project, we aim to comprehensively investigate the interplay between domain adaptation and PI in the context of medical image analysis. Drawing from the groundwork observed in [12], the focus of this project is to delve deeper into the optimal integration of diverse PI categories and their precision attributes. Through this endeavor, our objective is to enhance the robustness and effectiveness of image classifiers for medical applications.

This project focuses on investigating the interplay between domain adaptation and PI in the context of medical image analysis. By integrating both domain adaptation and PI, we aim to enhance the robustness and effectiveness of image classifiers for medical applications. The thesis follows a structured approach, beginning with a comprehensive background study to establish the context and identify research gaps in the field of robust medical imaging. It then delves into potential solutions, leading to the exploration of the "Unsupervised Domain Adaptation by Learning using Privileged Information" (DALUPI algorithms) and their applicability across diverse datasets.

Specifically, we aim to answer the following research questions:

1. Does the number of PI features provided impact the performance of the model? If so, in what way? Additionally, how does the correlation of the PI with the

output label impact the performance of the model?
2. To what extent does the granularity of PI, measured by specifics such as level of detail, relate to the performance of image classifiers in domains with notable distribution shifts?
3. How well can the DALUPI approach be extended to non-medical datasets, and what insights does its application in broader contexts provide? Can DALUPI consistently outperform existing domain adversarial models across various datasets?

The primary objective of the thesis is to evaluate the impact of leveraging PI on enhancing the robustness of the algorithm across varied datasets and modalities. Through a series of experiments, the thesis aims to address several research questions that remain unexplored in the context of DALUPI learning of image classifiers, particularly within the medical domain. The scope of this thesis encompasses the expansion and application of a developed algorithm aimed at evaluating its performance across diverse domains. The evaluation utilizes publicly available datasets, including the CelebA dataset, NIH Chestxray dataset, and the CheXpert dataset. Notably, the algorithm has not been assessed using proprietary or collected data sources.

In summary, the thesis provides a systematic exploration and expansion of DALUPI algorithms in the context of medical imaging, offering insights into their effectiveness, limitations, and future directions for research and development.

2
Background

This chapter serves as a foundational exploration of the theoretical aspects essential for the project's comprehension and execution. It is structured into subsections, each focusing on a major component relevant to the project's objectives.

2.1 Medical imaging

Medical imaging plays a vital role in examining the internal structures of the body to identify abnormalities. In cases of traumatic incidents, such as accidents, it is essential to detect haemorrhages promptly to prevent excessive blood loss and maintain the patient's stability. Timely identification of concussions or brain trauma is critical to prevent impairments in bodily functions. Moreover, medical imaging techniques are commonly used to detect fractures and tumours. These imaging modalities vary from basic X-ray scans to advanced techniques like MRI and Positron Emission Tomography (PET) scans.

2.1.1 Medical Imaging Modalities

Medical imaging modalities are broadly classified into anatomical and functional imaging. Anatomical imaging focusses primarily on visualising the structure and morphology of the body. These modalities include X-rays, Computed Tomography (CT), MRI imaging, and ultrasound imaging. While each modality operates on different physical principles, they all provide images that depict the structural features of the body.

X-rays are commonly utilized to detect fractures and abnormalities in bones due to their ability to penetrate tissues and produce images of dense structures. CT and MRI imaging offer more detailed views of internal organs and tissues. CT scans use X-rays from multiple angles to create cross-sectional images, allowing for the detection of clots, tumours, and other anomalies. MRI imaging, based on magnetic fields and radio waves, provides high-resolution images of soft tissues, making it particularly useful for diagnosing neurological disorders and identifying abnormalities in organs like the brain and heart.

Ultrasound imaging stands apart from other modalities as it does not involve exposure to ionizing radiation or radioactive materials. Instead, it utilizes sound waves to create images, making it safe for use during pregnancy and in paediatric patients.

Ultrasound imaging is commonly employed to assess blood flow, monitor foetal development during pregnancy, and evaluate the function of organs such as the heart and kidneys.

Anatomical imaging modalities [36] play a crucial role in the diagnosis of a wide range of medical conditions. Functional imaging modalities [37] specialize in visualizing the physiological processes and functions within the body. Among these, Single Photon Emission Computed Tomography (SPECT) and PET imaging are the prominent methods used today. Both modalities involve the administration of radioactive tracers to detect metabolic activity and functions of organs and tissues.

SPECT imaging involves the detection of gamma rays emitted by a radioactive tracer during its decay process, enabling the creation of three-dimensional images that illustrate the tracer's distribution within the body. This imaging modality is instrumental in assessing blood flow, identifying tissue damage, and diagnosing conditions like coronary artery disease and specific neurological disorders.

In contrast, PET imaging employs positron-emitting radioactive tracers that emit gamma ray pairs upon interaction with tissues. PET scanners then generate detailed images reflecting metabolic processes, such as glucose metabolism, within the body. Widely utilized in oncology for tumour detection, staging, and treatment response monitoring, as well as in neurology for diagnosing disorders like Alzheimer's disease and epilepsy, PET imaging provides crucial insights into organ and tissue functioning.

Both SPECT and PET imaging techniques furnish clinicians with indispensable information for diagnosing, planning treatments, and monitoring a diverse array of medical conditions. Despite their reliance on radioactive tracers, these modalities are deemed safe when administered judiciously and offer unparalleled capabilities in functional imaging.

While there are a multitude of imaging modalities to choose from for various instances, there are several challenges faced in interpreting these images. Challenges range from inherent noise and artefacts in the images to variations in image quality and resolution due to differences in equipment and imaging protocols. In addition, the anatomical variability among individuals and the complexity of pathological conditions further complicate the interpretation process.

2.2 Machine Learning in Medical Imaging

With the growing popularity of ML algorithms in recent years, they have become powerful tools for analysing medical images. ML techniques play a crucial role in tasks such as image segmentation, feature extraction, and classification, aiding clinicians in achieving accurate diagnosis and treatment planning.

Image Segmentation

Image segmentation is the process of partitioning an image into distinct regions or objects to simplify its analysis and interpretation. Medical images such as CT or MRI scans often contain a wealth of information. It is essential to focus on only the relevant information for better understanding and interpretation. In such cases, image segmentation is crucial for the extraction of important information [34]. ML algorithms, particularly deep learning models such as convolutional neural networks (CNN's) excel at image segmentation tasks [35]. By training on large datasets of annotated medical images, these algorithms can automatically identify and delineate structures and abnormalities within the images, such as tumors, organs, or blood vessels.

Feature Extraction

Once segmented, distinctive features or specific characteristics need to be extracted for analysis. These features range from shape, size, texture and intensity to spatial relationships between different structures. ML techniques such as Principle Component Analysis (PCA) and wavelets transforms can automatically detect informative features from medical images, reducing the manual effort required and potentially uncovering subtle patterns or biomarkers that may be indicative of specific diseases or conditions.

Classification Tasks

ML algorithms are often employed for classification tasks once the images are segmented and relevant features are discerned. The objective of the classification task is to assign a label or a diagnosis to each segmented region based on its extracted features. Various ML algorithms such as Support vector Machines (SVM) or Random Forest algorithms can be employed for the same. To enable automated diagnosis and assist clinicians in making informed decisions about patient care and treatment planning, these classifiers undergo training to recognize patterns and correlations within the derived features. Furthermore, ML models have the capacity to be updated and improved with fresh data on a regular basis, which eventually improves their accuracy and generalisation.

Machine learning algorithms can broadly be classified into three categories based on the availability of labelled data: supervised, semi-supervised and unsupervised learning. In *supervised machine learning* (SML) (A.1.1), models are trained using labeled data, pairing each input with its corresponding output. This method demands a considerable amount of labeled data to train effectively. *Unsupervised machine learning* (UML)(A.1.2), however, trains models on unlabeled data without explicit guidance. The objective is to uncover patterns, relationships, or structures within the data autonomously. *Semi-supervised machine learning* (SSML) (A.1.3) combines labeled and unlabeled data for training. This hybrid approach proves valuable when acquiring labeled data is costly or difficult. By leveraging the abundance of unlabeled data alongside limited labeled examples, SSML enhances model performance.

The following subsections primarily discusses the range of algorithms utilized in the experiments conducted for this thesis. This includes various types of Convolutional Neural Networks, and commonly used domain adaptation algorithms to asses the effectiveness of the proposed algorithm.

2.2.1 Input, Output and Loss

Typically, ML models have one or more inputs, denoted as x. These inputs can vary widely, including text, audio, sensor data, tabular data, images, etc. The model then predicts an output, denoted as y. The output y represents the target variable or variables that the model aims to predict based on the input x. This is depicted as a simplified block diagram in fig 2.1 For instance, in regression tasks, y is a continuous value, while in classification tasks, y consists of a set of discrete classes or categories.

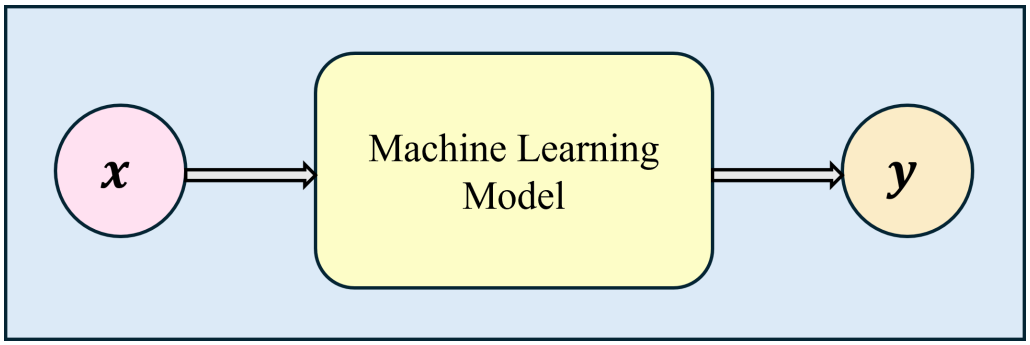


Figure 2.1: Simplified block diagram of a ML model with input x and output y

The primary goal in ML is to train models to make accurate predictions or classifications based on input data x. This is achieved by optimizing model parameters to minimize a predefined loss function, which quantifies the discrepancy between predicted outputs $\hat{y}$ and actual labels y.

Mathematically, the loss function $\mathcal{L}$ can be defined as the average discrepancy between predicted and actual values over a dataset of size N:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\hat{y}_i, y_i) \tag{2.1}$$

Where:

- θ represents the parameters of the model.
- $\hat{y}_i$ denotes the predicted output for the ith sample.
- y_i is the actual label for the ith sample.
- $\mathcal{L}$ is a loss function that computes the discrepancy between predicted and actual values for a single sample.

2.2.2 Convolutional Neural Network (CNN)

A CNN is a type of deep learning algorithm (an algorithm which can learn from its own errors [62]), which is particularly suited for image recognition and processing tasks [63]. The main layers seen in a CNN are given in Table 2.1 [64] [65].

Table 2.1: Overview of Different Layer Types in CNNs

Layer Type	Description
Convolutional Layers	Utilize convolutional operations to detect features like edges and textures in input images. Maintain spatial relationships between pixels.
Pooling Layers	Reduce spatial dimensions of input, lowering computational complexity. Common operations include max pooling, selecting the maximum value from groups of adjacent pixels.
Fully Connected Layers	Make predictions based on high-level features learned from previous layers. Establish connections between every neuron in one layer to every neuron in the next layer.

The convolutional layer is a fundamental component of a CNN. This layer applies filters to the input image, extracting essential features such as edge detection, textures, and shapes. The output from this layer is subsequently passed to the pooling layers, which serve to down-sample the feature maps, reducing spatial dimensions while preserving the significant features extracted in the previous layer. The output of the pooling layer then undergoes one or more fully connected layers, which are responsible for predicting or classifying the image based on the designated task. A typical CNN is depicted in fig 2.2

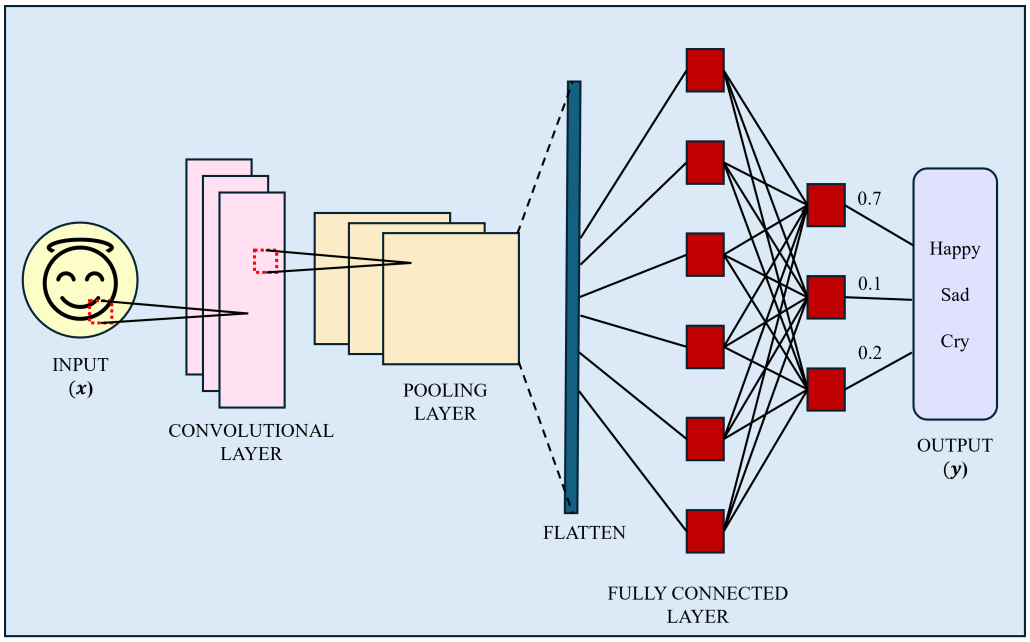


Figure 2.2: Generalised CNN Architecture

The neural network training process involves adjusting the weights of the network to minimize a predefined loss function [73]. This process is known as back-propagation [72], and it relies on gradient descent [74] to iteratively update the weights. The gradient descent algorithm computes the gradient of the loss function with respect to the weights and adjusts the weights in the opposite direction of the gradient to minimize the loss [75]. The weights are updated according to the following equation:

$$w_{t+1} = w_t - \alpha \frac{\partial \mathcal{L}}{\partial w} \tag{2.2}$$

Where:

- w_{t+1} is the updated weight at iteration $t + 1$.
- w_t is the current weight at iteration t.
- α is the learning rate, which controls the step size of the weight updates.
- $\mathcal{L}$ is the loss function.
- $\frac{\partial \mathcal{L}}{\partial w}$ is the gradient of the loss function with respect to the weights.

The choice of loss function depends on the task at hand. For classification tasks, common loss functions include cross-entropy loss and hinge loss. For regression tasks, mean squared error (MSE) and mean absolute error (MAE) are commonly used.

The versatility of CNNs spans multiple applications, such as image classification, object detection, segmentation, and generation [66]. For object detection tasks, bounding boxes are commonly employed. These boxes identify and encapsulate the targeted object, allowing for targeted attention during subsequent processing and analysis. Often, multiple boxes may be used to represent different objects within

the same image, as the total number of objects is not predetermined. Conventional CNNs struggle in these scenarios due to the variability in the output layer's length, which corresponds to the unpredictable number of object occurrences. Under such circumstances, algorithms like R-CNN [76] and YOLO[102] prove to be more adept.

2.2.2.1 Residual Network (ResNet)

ResNets [103], short for Residual Networks, are a type of deep learning model commonly used in computer vision tasks, particularly for image classification. They are specialized CNNs capable of supporting hundreds or even thousands of layers. In traditional deep neural networks, as more layers are added, the gradient during backpropagation tends to diminish, leading to the problem of vanishing gradients. This phenomenon can degrade the performance of the model as it becomes increasingly difficult to train deeper networks effectively.

ResNets address this issue by introducing "skip connections", or shortcuts, that allow the network to bypass certain layers. This enables the gradient to flow more directly through the network during training, mitigating the vanishing gradient problem and facilitating the training of extremely deep neural networks. As a result, ResNets can effectively leverage a large number of layers while maintaining or even improving performance compared to shallower networks. The architecture of a typical ResNet is shown in fig 2.3.

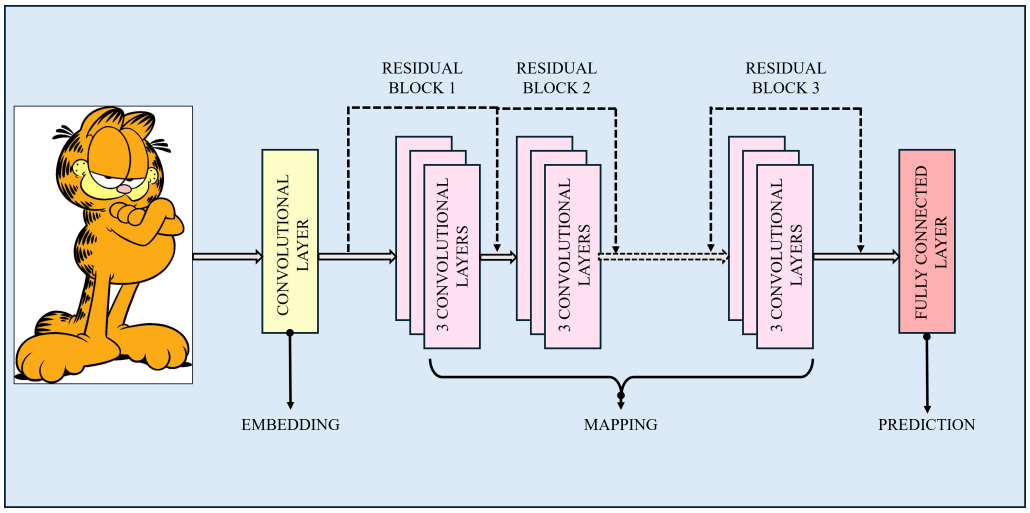


Figure 2.3: Generalised ResNet Architecture

In conventional CNNs, the convolutional layers are stacked with batch normalization and non-linear activation layers such as ReLu between them. This model is known to work well with a small number of convolutional layers such as in the VGG16 [69] or VGG19 [70] [71] architectures. The ResNet architecture implements "residual blocks". These blocks add an intermediate input to the output of a series of convolutional blocks [76].

The residual block can be expressed as:

$$\mathrm{H}(x) = F(x) + x \tag{2.3}$$

where x is an input to the residual block and output from the previous layer, $H(x)$ is the output of the residual block and $F(x)$ is a function representing the residual mapping learned by the network [77] [78] [79].

2.2.2.2 Region-based Convolutional Neural Network (R-CNN)

To address the issue of selecting numerous regions, Girshick et al. introduced a technique to extract 2000 regions from an image [67]. Referred to as region proposals, these selected areas significantly reduce the overwhelming task of classifying numerous regions to just 2000. These region proposals are then processed by a CNN, which extracts features and relays them to an SVM through a dense output layer for object presence classification within the region proposal. Furthermore, the algorithm predicts offset values to enhance the accuracy of the bounding box as seen in fig 2.4. For instance, if the algorithm predicted a bird's presence in a region proposal, it might only capture half of the bird. Thus, the offset values are crucial for adjusting the bounding box to fully encompass the region proposal.

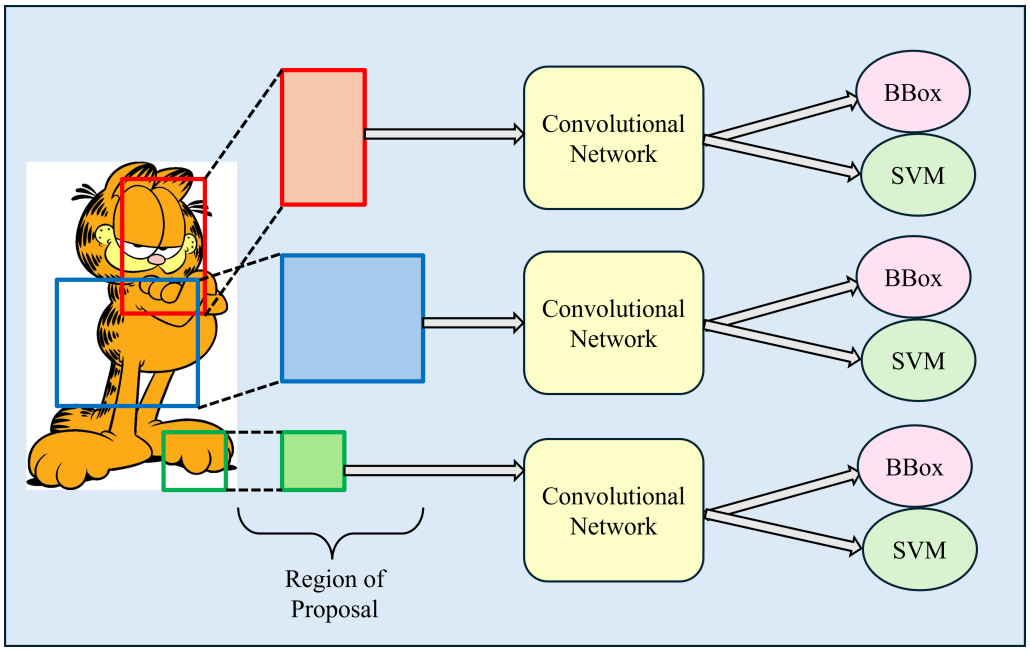


Figure 2.4: Generalised R-CNN Architecture

While R-CNNs address the challenge of selecting numerous regions, they still pose computational challenges. Training the network can be time-consuming due to the need to classify a large number of images, typically around 2000. Additionally, the selective algorithm used is fixed, meaning that no learning occurs during this stage. This increases the possibility of generating poor region proposals, which can impact the overall performance of the model [68].

2.2.2.3 Fast Region-based Convolutional Neural Network (Fast R-CNN)

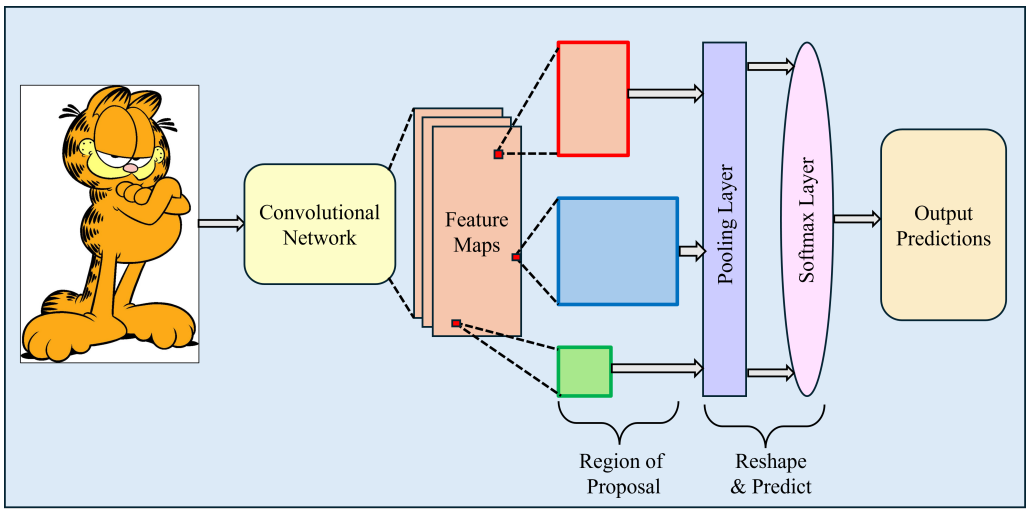


Figure 2.5: Generalised Fast R-CNN Architecture

Girshick et al. introduced a refined version of the R-CNN algorithm to overcome its limitations. The Fast R-CNN model retains the general architecture of R-CNN but introduces significant improvements. Rather than directly feeding proposed regions to the CNN, the entire input image is passed through the CNN to generate a convolutional feature map. This feature map is then utilized to identify regions of interest (RoIs), which are subsequently warped into squares. A ROI pooling layer is applied to reshape these regions into a fixed size, enabling them to be fed into a fully connected layer. Finally, a softmax layer is employed to predict both the class of the proposed region and the offset values for the bounding box as seen in fig 2.5.

2.2.3 Challenges Faced in Imaging

The efficacy of medical imaging in clinical practice is significantly impacted by several key challenges. Image quality often suffers from noise, artifacts, and variations due to differences in imaging equipment. Anatomical variability further complicates matters by hindering the establishment of standardized reference points for interpretation. Moreover, the complexity and diversity of diseases and pathological conditions present significant challenges, necessitating advanced imaging techniques and expert interpretation to identify subtle abnormalities and differentiate between similar pathologies.

In photography or imaging, a notable absence of rigid regulations results in a diverse array of methodologies employed to capture images. These methods exhibit variations in resolution, size, and colour depth, tailored to meet specific contextual demands. Similarly, the field of medical imaging lacks standardized protocols, with the approach to image acquisition dictated by a multitude of factors including the unique anatomical features of each individual and the technical capabilities of the imaging devices, such as X-ray machines.

Despite certain imaging machines boasting superior image clarity, subtle discrepancies in the processing algorithms employed can give rise to variations in the resultant images. This phenomenon leads to the emergence of distinct "domains" within the same modality, characterized by subtle differences in image characteristics. Adapting algorithms to effectively handle this diversity poses a considerable challenge. Domain adaptation techniques are being explored to address these challenges, aiming to develop robust algorithms capable of accommodating the diverse imaging characteristics encountered in clinical practice.

2.3 Domain Adaptation Challenges

Domain adaptation, a prominent challenge in the field of machine learning, involves the task of adjusting a model trained on data from one domain (the source domain) to perform effectively on data from a different domain (the target domain), which possesses distinct characteristics. In the context of medical imaging, domain adaptation may entail training a machine learning model on images sourced from one hospital or medical institution and subsequently adapting it to accurately process images obtained from a different institution, thereby ensuring robust performance across diverse data sources.

Domains in medical imaging can encompass various factors such as patient demographics, imaging protocols, equipment specifications, and environmental conditions, all of which contribute to variations in the underlying data distribution. The process of domain adaptation can be classified along two main axes. Firstly, based on the availability of labelled data, supervised domain adaptation (SDA), Semi-supervised domain adaptation (S-SDA) and unsupervised domain adaptation (UDA).

Secondly, domain adaptation techniques can be classified based on the perspective of adaptation, including model-centric, data-centric, and hybrid strategies. In model-centric adaptation, the focus is on adapting the model architecture or parameters to align features between the source and target domains. This involves fine-tuning pre-trained models or incorporating domain-specific adaptation layers in the model. In contrast, data-centric adaptation aims to minimise domain discrepancy by adjusting the data distributions or representations. This includes methods like domain-specific data augmentation and domain adversarial training.

2.3.1 Domain-Adversarial Neural Network

In 2016, Ganin et al. [32] proposed a new learning approach for domain adaptation. The theory of domain adaptation suggests that, in order to accomplish successful domain transfer, predictions need to be built on features that are unable to distinguish between the training (source) and test (target) domains. The paper introduced Domain-Adversarial Neural Network (DANN), an algorithms combining the techniques of both representation learning (deep feature learning) and UDA. The end to end algorithm develops a model with features to solve a task for both the source

and target domains, given labelled samples from the source and unlabelled samples from the target distribution [33]
As seen in fig 2.6, the DANN framework encompasses three primary components: a feature extractor, a domain classifier, and a task-specific classifier [80]. DANN leverages a domain classifier to motivate the network to learn features that are invariant to domain variations (domain-invariant features). The feature extractor, denoted as G, learns to extract informative representations from the input data, including PI characteristics. These representations, denoted as h_f, are then used by both the domain classifier and the discriminative classifier. The domain classifier, denoted as D_{domain}, endeavors to distinguish between the source and target domains during training. Meanwhile, the discriminative classifier, denoted as D_{class}, is trained to predict task-specific class labels based on the extracted features h_f.

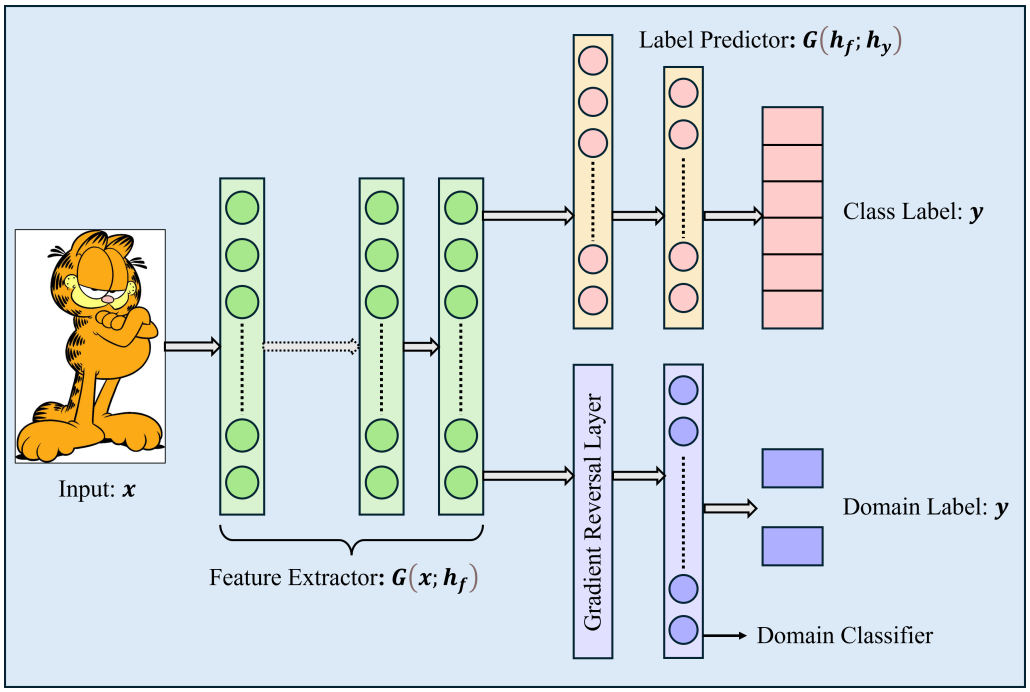


Figure 2.6: Generalised DANN Architecture

During training, the network is optimized using an adversarial training objective, which prompts the feature extractor to generate domain-invariant features [81]. This objective aims to minimize the discrepancy between domains, thereby facilitating robust domain adaptation. The adversarial training objective is formulated as follows:

$$\min_{G} \max_{D_{\text{domain}}} \mathcal{L}_{\text{task}}(G, D_{\text{class}}) - \lambda \mathcal{L}_{\text{domain}}(G, D_{\text{domain}}) \tag{2.4}$$

Here, $\mathcal{L}_{\text{task}}$ represents the loss function for the task-specific classification, while $\mathcal{L}_{\text{domain}}$ is the loss function for domain classification. The trade-off parameter λ controls the importance of the domain adaptation objective relative to the task-specific objective.

This equation represents the minimax optimization problem where the feature extractor G and the domain classifier D_{domain} are optimized simultaneously, aiming to minimize the task-specific loss while maximizing the domain classification loss with a trade-off controlled by the parameter λ.

$$D_{\text{class}}(x) = P(\text{class}|x) \tag{2.5}$$
$$D_{\text{domain}}(x) = P(\text{source}|x) \tag{2.6}$$
$$G(x; h_f) = h_f \tag{2.7}$$
$$G(h_f; h_y) = \hat{y} \tag{2.8}$$

Equation 2.5 represents the discriminative classifier D_{class}, which predicts the class label based on the input x. Equation 2.6 denotes the domain classifier D_{domain}, which determines whether the input x belongs to the source domain or the target domain. The feature extractor G extracts meaningful features from the input data, as shown in Equation 2.7. Finally, Equation 2.8 illustrates the label predictor, which produces the predicted class label $\hat{y}$ based on the extracted features h_f [82].

2.4 Privileged Information

As humans, we often rely on subtle cues or characteristics to make distinctions that seem obvious to us, such as distinguishing between men and women based on attributes such as makeup, long hair or beard. However, these distinctions may not be as apparent to machine learning models, as they have not been explicitly learnt. To address this, we can provide the model with additional information, known as 'privileged information', during training.

Introduced by Vapnik and Vashist [4], privileged information refers to extra data or knowledge that is available during the training phase but may not be accessible during testing or when making predictions. By incorporating this additional insight into the training process, the model can learn to better differentiate between classes or make more accurate predictions. For instance, in medical diagnosis, privileged information could include detailed patient history, lab test results, or other diagnostic information that may not always be available at the time of prediction. Table. 2.4.1 lists the various kinds of PI seen in medical imaging. By leveraging this privileged information during training, machine learning models can improve their performance and effectiveness in various applications.

Table 2.4.1: Types of PI seen in Medical Imaging

PI Feature	Example in medical Imaging
Bounding box	Region of interest in an image
Segmented image	Object segmentation maps
Time series	Temporal data sequences
Semantic concepts	Keywords, categories, labels
Expert annotations	Keywords, notes from a domain specialist
Historical data	Previous records or trends

In simpler terms, UDA is similar to teaching a model to understand a new environment without explicitly telling it what each thing is called [83, 84]. Instead, the model learns to recognize patterns and similarities between the source and target domains, allowing it to make accurate predictions even in the new, unfamiliar setting. This is particularly useful in scenarios where collecting labelled data for the target domain is difficult or expensive. By leveraging the similarities between the source and target domains, unsupervised domain adaptation enables the model to generalize its learning to new, unseen data more effectively.

Acquiring labelled medical image data poses significant challenges due to the labour-intensive nature of the annotation process [85, 86]. Medical professionals, who are responsible for this task, often face time constraints that hinder their ability to meticulously annotate and label data that can be used for ML algorithms. In many cases, relevant information is discussed verbally and documented in reports, rather than being directly annotated on the images themselves. However, ML relying on annotations derived from reports is less reliable compared to direct annotations, as it introduces potential inaccuracies and inconsistencies in the labelled data.

Medical image classification becomes even more challenging when the target domain lacks sufficient annotated data, which is often the case in medical imaging due to factors such as privacy concerns, data scarcity, and the need for specialized expertise. UDA emerges as a promising solution to address these challenges. UDA techniques enable the transfer of knowledge from a labelled source domain to an unlabelled target domain, allowing machine learning models to generalize effectively even in the absence of labelled data in the target domain. This is particularly advantageous in medical imaging scenarios where labelled samples in the target domain are sparse or non-existent. By leveraging the similarities between the source and target domains, UDA facilitates the adaptation of pre-trained models to the target domain, thereby enhancing their performance and robustness.

One of the key advantages of UDA is its ability to overcome the limitations of supervised learning approaches, which heavily rely on the availability of labelled data[88]. Instead of relying solely on labelled samples, UDA leverages the underlying data distribution in both the source and target domains to learn domain-invariant representations. This enables the model to effectively generalize to unseen data in the target domain, even when labelled samples are scarce or unavailable [87].

Moreover, UDA approaches may be able to reduce the effects of domain shift, a phenomenon in which variations in the data distribution between the source and target domains negatively impact model performance. By aligning the feature distributions across domains, UDA ensures that the model can accurately capture relevant patterns and information in the target domain, thereby improving its overall performance in medical image classification tasks. However, employing UDA as the primary methodology necessitates substantial assumptions for successful model performance. Addressing these challenges, Breitholtz et al. propose a novel approach called "Unsupervised Domain Adaptation by Learning using Privileged Information" (DALUPI), which seeks to circumvent these assumptions. In the following chapter, we will delve deeper into the methodology employed to evaluate the effectiveness of DALUPI in improving model performance and robustness in medical image classification tasks.

3
Methods

This chapter outlines the methods, tools, and techniques utilized in the project, following a structured framework from data acquisition and pre-processing to model development, training, evaluation, and validation.

3.1 Softwares and Tools

The DALUPI algorithm was implemented in Python, a high-level, object-oriented programming language [40]. Python contains several built-in libraries such as NumPy, Seaborn, and scikit-learn, among many others. These libraries comprise a collection of code that enhances the efficiency of coding in Python [49]. Python is widely favored for data analysis due to its extensive arsenal of tools for data manipulation, visualization, and machine learning model training. The experiments were conducted on the Alvis cluster, a resource provided by NAISS (National Academic Infrastructure for Supercomputing in Sweden). Alvis is tailored to support research involving Artificial Intelligence techniques [41], offering specialized infrastructure for machine learning and artificial intelligence tasks. All data, scripts, and jobs were stored and executed on Alvis.

3.2 DALUPI Method

To effectively use PI for addressing domain adaptation challenges, we aim to integrate domain adaptation and task-specific supervised learning techniques. Through this, we aim to exploit the informative nature of PI to effectually mitigate domain shift effects and enhance the prediction accuracy and model adaptability. We align our selection of supervised learning methods with the specific demands of each task, ensuring that we optimize the utilization of PI to cater to the distinctive characteristics and objectives of each task. For a deeper understanding of the core algorithm proposed in this study, this subsection looks at the DALUPI method in more detail.

As touched on briefly in the in the previous section, the Unsupervised Domain Adaptation using Privileged Information or DALUPI approach was proposed by Breitholtz et al. [12]. DALUPI is a novel machine learning approach that integrates unsupervised domain adaptation with PI. It leverages PI, which is accessible only during training, to enhance model performance in adapting to target domains with distinct data distributions, particularly in scenarios with limited labelled data. The

DALUPI algorithm works under three assumptions, that of covariate shift, domain overlap and sufficiency of PI.

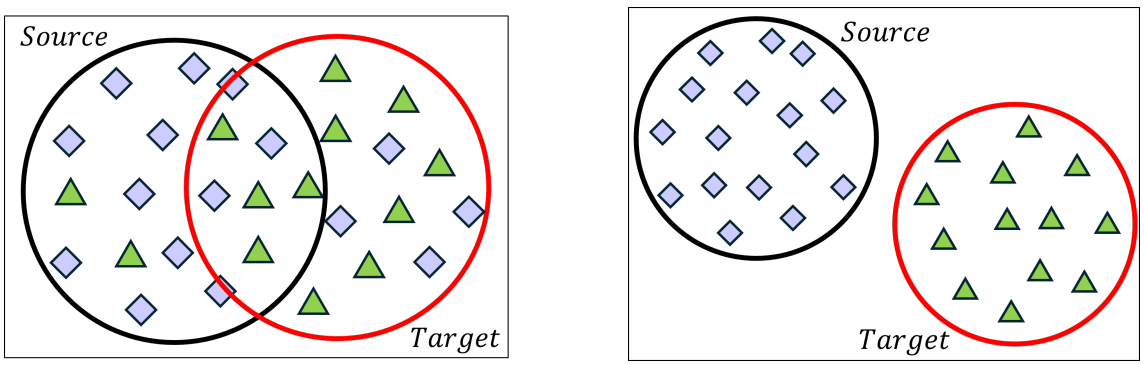
Covariate shift in machine learning refers to the situation where the conditional probability distribution of the input data ($P(x|y)$) differs between the training dataset (denoted as $P_s(x|y)$) and the testing dataset (denoted as $P_t(x|y)$), which can be expressed as:

$$P_s(x) \neq P_t(x)$$

This distributional mismatch can lead to reduced model performance when applying the model trained on the source domain (P_s) to the target domain (P_t) [89].

Domain overlap, in the context of machine learning, occurs when there is a significant intersection or similarity between the input data distributions ($P(x)$) of two or more domains, typically the source domain ($P_s(x)$) and the target domain ($P_t(x)$) as seen in fig 3.1. Mathematically, domain overlap can be expressed as:

$$P_s(x) \cap P_t(x) \neq \emptyset.$$



(a) Domain Overlap **(b)** No Domain Overlap

Figure 3.1: The presence and absense of domain adaptation between source and target domains

In this scenario, some input data points are shared or similar between the domains, which may affect the model's performance during domain adaptation tasks.

The *sufficiency of PI* refers to the extent to which additional data or information provided during the training phase (PI, denoted as w) contains relevant and informative content that contributes to improving the model's performance on a specific task in the target domain $\mathcal{T}$. It can be mathematically expressed as:

$$I(x, y|w) = 0$$

where $I(x, y|w)$ represents the conditional mutual information between variables x and y, given the privileged information w. This condition indicates that the PI w contains all the necessary information required for accurate predictions (y) in

the target domain, given the input data (x). When $I(x, y|w) = 0$, w is considered sufficient for the task, as it eliminates the need for additional information to make accurate predictions. This definition holds true only under the assumption that there is conditional independence between x, y, and w such that w fully captures the relevant information for the task.

Keeping these three assumptions in mind, the DALUPI approach can be implemented in two ways: Direct approach and the two-staged approach. Applying the above assumptions to PI w enables us to identify the target risk $R_{\mathcal{T}}(h)$ for models $h \in \mathcal{H}$ that don't use w as input.

$$R_{\mathcal{T}}(h) = \sum_{x} \mathcal{T}(x) \sum_{w} \mathcal{T}(w \mid x) \sum_{y} \mathcal{S}(y \mid w) \mathcal{L}(h(x), y) \tag{3.1}$$

For the squared loss ($\mathcal{L}$), the minimizer of $R_{\mathcal{T}}$ is given by:

$$h^*_{\mathcal{T}}(x) = \sum_{w} \mathcal{T}(w \mid x) \mathbb{E}_{\mathcal{S}}[Y \mid w] \tag{3.2}$$

where Y denotes the target variable associated with input x.

3.2.1 Two-stage approach

This implementation approach entails estimating the inter-dependencies among input data, PI, and target outcomes and subsequently amalgamating these associations to enhance predictive accuracy.

Step 1: Modeling PI
In the context of the DALUPI algorithm, the first crucial step involves modeling PI. This step is fundamental for enhancing the depth of information available for predictions. We represent the relationship between PI and the input data (x) as $\mathcal{T}(w|x)$. This modeling enables the system to understand how PI varies under different input conditions, providing valuable insights for subsequent predictions.

Step 2: Outcome Modeling
Simultaneously, the second step focuses on modeling the outcome (y) based on the provided PI. We denote this approximation as $\hat{g}(w)$, which approximates the expected outcome $\mathbb{E}_{\mathcal{S}}[y|w]$. By constructing this outcome model, we gain a better understanding of how outcomes relate to the PI. This step significantly contributes to more informed predictions.

Step 3: Final Integration
The final combination of these insights is crucial for effective prediction, especially when PI is determined as a deterministic function of the input (x). In this scenario, we estimate a mapping function $f : \mathcal{X} \to \mathcal{W}$ through regression ($\hat{f}$). The estimated mapping $\hat{f}$ is used to predict PI based on input data ($\hat{f}(x)$). The final prediction ($\hat{h}$) integrates information from both the input (x) and the estimated PI ($\hat{f}(x)$). This integrated approach approximates the expected outcome as $\mathbb{E}_{\mathcal{S}}[y|\hat{f}(x)]$. The

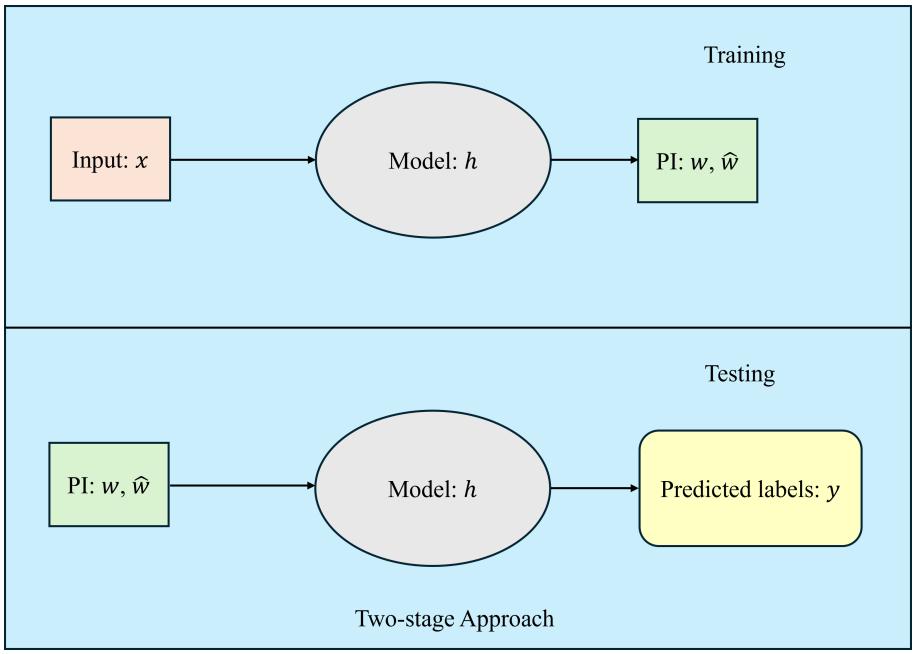


Figure 3.2: Block diagram depicting the two-stage approach of DALUPI

optimization of functions $\hat{f}$ and $\hat{g}$ is achieved through empirical risk minimization, enabling the most accurate predictions by leveraging the interplay between input, PI, and outcome.

This two-step approach provides a comprehensive framework for effectively incorporating PI into machine learning models, enhancing their predictive capabilities. Figure.3.2 displays a simplified block diagram of this implementation of the algorithm.

3.3 Experiment Design

To assess the influence of different PI's on diverse datasets, two separate experiments were devised. Each experiment involved the deployment of a version of the DALUPI algorithm. The datasets and types of PI's varied between experiments, as did the types of tests conducted. The framework followed for each experiment is shown in Figure.3.3

Table.3.3.1 describes the various kinds of PIs and datasets available for studying. While there are several datasets that contain information that can be used as PI, some of these are more complex to use than others. These complexities stem from the necessity to handpick relevant information that can used as PI such as annotating data, extracting relevant patient history or keywords, and inferring trends observed from clinical notes, among others. For simplification of these experiments, the selected PIs included semantic concepts, specifically attributes, bounding boxes, and segmentation [90–99].

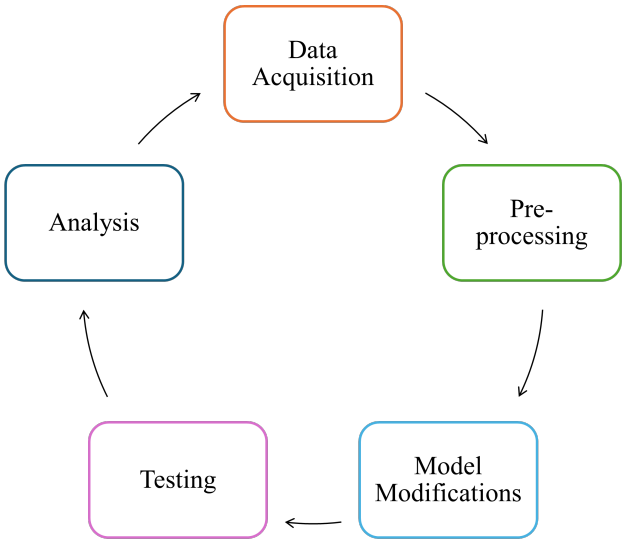


Figure 3.3: Experiment Design

Table 3.3.1: Features, Datasets, and Descriptions

Feature	Datasets	Description
Bounding box	CheXpert, NIH Chest X-ray	Chest X-ray datasets containing annotations or bounding boxes around specific anatomical regions.
Segmented image	IXI, BRaTS,	Medical image datasets with segmentation annotations indicating distinct structures.
Time series	MIMIC-II Notes	Temporal data sequences such as clinical notes and reports from the MIMIC-II database.
Semantic concepts	ADNI, MIMIC, CelebA	Datasets with semantic labels, keywords, or categories associated with medical images.
Expert annotations	ADNI, BRaTS	Annotations provided by domain experts, enhancing image understanding.
Historical data	MIMIC-II Notes	Historical patient records and trends from clinical notes in the MIMIC-II dataset.

3.4 Experiment I: Binary Attributes

In the first experiment, attributes are utilized as the PI, and their impact is assessed on the CelebA dataset. The task here is binary classification of the data into labels male or female. In this case attributes refer to adjectives used to describe the image such as ‘heavy makeup’, ‘bald’, ‘ moustache’ and so on. The overall experiment is

designed to assess the impact of the number of attributes on the prediction accuracy. Additionally, variations in attribute combinations are explored to detect any significant changes in efficiency. It is noted that the dataset is unbalanced, with differing numbers of male and female samples. No oversampling is performed, and the raw data is used as is.

3.4.1 Hypotheses: Attributes

Two hypotheses are tested:

- *Hypothesis 1:* Given the varying correlations between attributes and labels, it is hypothesized that after a certain number of attributes used as PI, the model output will exhibit stabilization, meaning that further addition of attributes as PI will not significantly improve the model's performance or prediction accuracy.

- *Hypothesis 2:* It is proposed that the sequence in which attributes are integrated into the model will influence the number of attributes required to attain a stable output accuracy.

3.4.2 CelebA Dataset

The CelebFaces Attributes Dataset (CelebA) [39] contains over 200,000 images of celebrities, each annotated with 40 attributes. All the images in this dataset created by Liu et al. are obtained from the internet. The dataset covers various facial poses and background complexities, offering extensive diversity and comprehensive annotations. It includes 10,177 distinct identities, 202,599 facial images, and annotations for 5 landmark locations and 40 binary attributes per image. CelebA is widely used in computer vision tasks such as face attribute recognition, face recognition, face detection, landmark localization, and face editing & synthesis. The dataset can be accessed directly from the website and downloaded [50]. The use of these images are restricted to non-commercial research and educational purposes.
Figure 3.4 displays sample images extracted from the CelebA dataset, each accompanied by its corresponding attributes. The complete list of attributes accessed from the dataset is given in Table.A.3.1. The last column of the table indicates the label assigned by a human after reviewing the attribute. Human descriptions were employed as a preliminary test to evaluate the model's predictive capability. This approach served to highlight discrepancies between model predictions and human intuition, as in real-world scenarios, certain attributes, such as a moustache on a female, are improbable. Additionally, this column facilitated the assessment of the correlation between each attribute and the labels assigned by humans.

3.4.3 Attribute-Label Correlation

As seen in Figure 3.4, several images correlate with the same attribute. In this experiment, the direct implementation of DALUPI was employed. The impact of

Figure 3.4: Sample images from CelebA with corresponding attributes. Source: [39]

these attributes was evaluated through two testing methods. First, the correlation of each attribute with the label was calculated. Certain attributes were expected to show a stronger correlation with the "Male" label, while others would correlate more strongly with the "Female" label.

After obtaining the correlations, the attributes were divided into two lists. List-1, Attributes Ranked by Correlation with Male Label (ACML) is arranged the attributes in descending order of correlation with the "Male" label, with the first entry having the highest correlation and the last entry having the lowest correlation with the label. List-2, Attributes Ranked by Overall Correlation (AOC) organised the attributes from highest to lowest correlation, regardless of the label. Table.3.4.2 and table.3.4.3 contain the first five attributes on ACML and AOC. For this model, "Wearing a hat" is set as the in-domain selector, and the "Male" label is not taken into consideration. Therefore, the target and source domains are defined by people wearing ($\mathcal{T}$) and not wearing (S) a hat.

Table 3.4.2: Attributes and Correlation to Label: Male

Serial	Attribute	Correlation to Label: Male
1	5 o'clock Shadow	0.416
2	Big Nose	0.373
3	Wearing Necktie	0.329
4	Bags Under Eyes	0.302
5	Goatee	0.306

Table 3.4.3: Attributes and Overall Correlation

Serial	Attribute	Overall Correlation
1	Heavy Makeup	0.666
2	No Beard	0.521
3	5 o'clock shadow	0.416
4	Arched Eyebrows	0.407
5	Attractive	0.400

The tables display attributes in different orders based on their correlation with labels and overall correlation. These lists will be used to evaluate how attributes impact the DALUPI model when utilized as PI.

3.4.4 Image Preprocessing

The images in the were converted to Red-Green-Blue (RGB) format, each with dimensions of 64x64 pixels. These dimensions align with standard practices in deep learning. Training parameters, including the number of epochs (100) and batch size (32), adhered to established conventions for model training. During the fine-tuning phase, a learning rate of 1.0e-5 and a maximum of 20 epochs were utilized to balance accuracy and computation time. Data processing allocated splits among labeled, unlabeled, validation, and test sets.

Table 3.4.4: Hyperparameters

Hyperparameter	Value
batch_size	[16, 32, 64]
learning rate	[1.0e-5, 1.0e-4, 1.0e-3]
optimizer___weight_decay	[1.0e-4, 1.0e-3]
callbacks___lr_scheduler___step_size	[15, 30, 100]
experiment	'celeb'
seed	0 to 4
image_w	64
image_h	64
image_c	3
n_epochs	100
batch_size	32
n_classes	2

Table 3.4.4 outlines the range of hyperparameters utilized for refining the model. To ensure reproducibility, different combinations of these hyperparameter settings were tested, with each combination seeded from 0 to 4. In total, 10 sweeps were executed with various combinations. The optimal tuning setting was selected based on achieving the highest accuracy.

3.4.5 Model Configurations

The experiment was designed for binary classification, with two specified classes: 'Male' and 'Female'. To assess the impact of the attributes on the label prediction, each iteration of the experiment implemented a different configuration of the model. First, the ACML list was utilised. The experiment used only first attribute on the list as PI. The corresponding results were recorded. Subsequently, a systematic approach was adopted, incorporating additional attributes in sequential iterations. Each experiment was carried out over 10 epochs with five different hyperparameter sweeps to test the model's behaviour under various scenarios. This iterative process continued until all 37 attributes were integrated as PI in the model. All the results were stored for evaluation. The same experimental procedure was repeated using the AOC list, and the results were saved for further analysis.

3.5 Experiment II: Bounding boxes

The seconds experiment designed implemented the two-stage DALUPI algorithm on two datasets: NIH Chest X-ray [42] and the Stanford CheXpert datasets [43]. The PI focused on here are bounding boxes. The experiment is designed to evaluate the influence of the size of the bounding box on the models prediction accuracy. This experiment was also designed to test the robustness of the algorithm on image. Additionally, the experiment serves to evaluate the algorithm's robustness across images sourced from two distinct domains.

3.5.1 Hypotheses: Bounding box

Three specific hypotheses are tested:

- *Hypothesis 3:*
  It is hypothesized that if the bounding box is excessively large (e.g., covering the entire image), DALUPI should exhibit behavior akin to an algorithm designed without any PI.
- *Hypothesis 4:*
  Conversely, it is hypothesized that if the bounding box is extremely small (almost negligible in size), DALUPI's performance should remain unaffected, resembling that of any other algorithm without PI.
- *Hypothesis 5:*
  It is postulated that shifting the position of the predicted bounding box will lead to a reduction in model accuracy proportional to the degree of shift.

3.5.2 Dataset 1: NIH Chest X-Ray

The NIH Chest X-ray is extracted from the clinical PACS [44] database at the National Institutes of Health Clinical Centre. It consists of approximately 60% of all the frontal chest x-rays in the hospital. The database contains images pertaining to

fourteen disease categories including Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural thickening, Cardiomegaly, Nodule, Mass and Hernia Figure.3.5 contains eight examples of these common thorax diseases. For this experiment, Atelectasis (AT), Cardiomegaly (CM) and Pleural Effusion (PE) are considered.

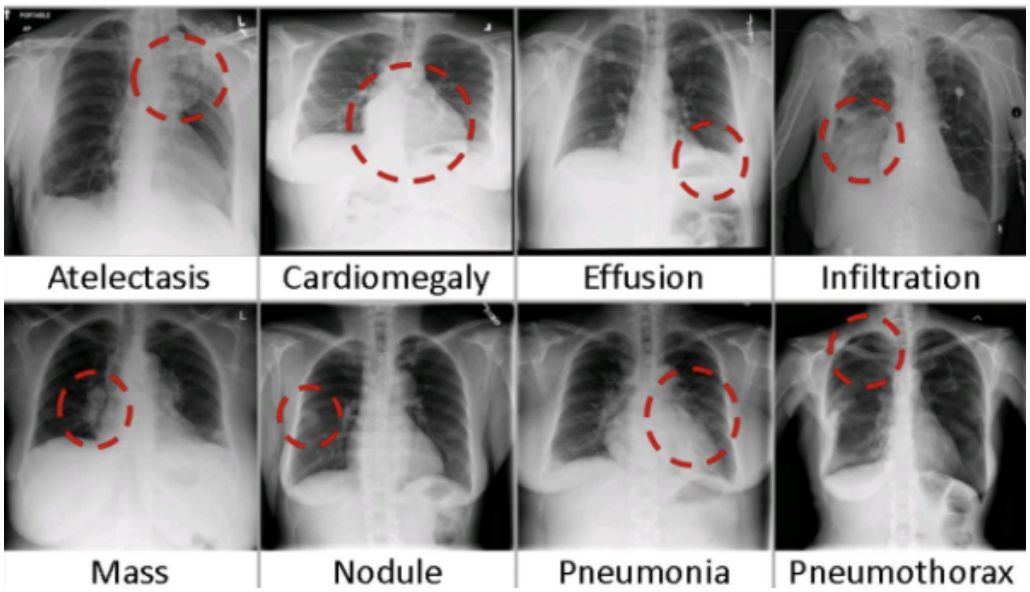


Figure 3.5: Sample images from NIH Chest X-ray showing common thorax diseases as taken directly from the data source. Source: [42]

The dataset comprises 112,120 images, each with a resolution of 1024x1024 pixels. Additionally, a list of bounding boxes is available for approximately 1000 images, along with two data split files for training, validation, and testing purposes. The labels are extracted using Natural Language Processing (NLP) and are estimated to have an accuracy exceeding 90%.

3.5.3 Dataset 2: CheXpert

Chexpert is a dataset curated by the Stanford ML group [43]. It is a large public dataset for chest radiographs, consisting of 224,316 images collected from 65,240 patients over a period of 15 years from the Stanford hospital. To label the data, the images were tested for the presence of fourteen observations based on their prevelance in their corresponding clinical reports. An automatic rule-based labeler was designed and used to label the dataset. The fourteen observations were Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pneumonia, Pneumothorax, Pleural Other, Support Devices and No Finding.

3.5.4 Bounding Box Estimation

To create a model applicable to images from both datasets, it was necessary to identify common labels. Upon comparing the disease categories across the datasets

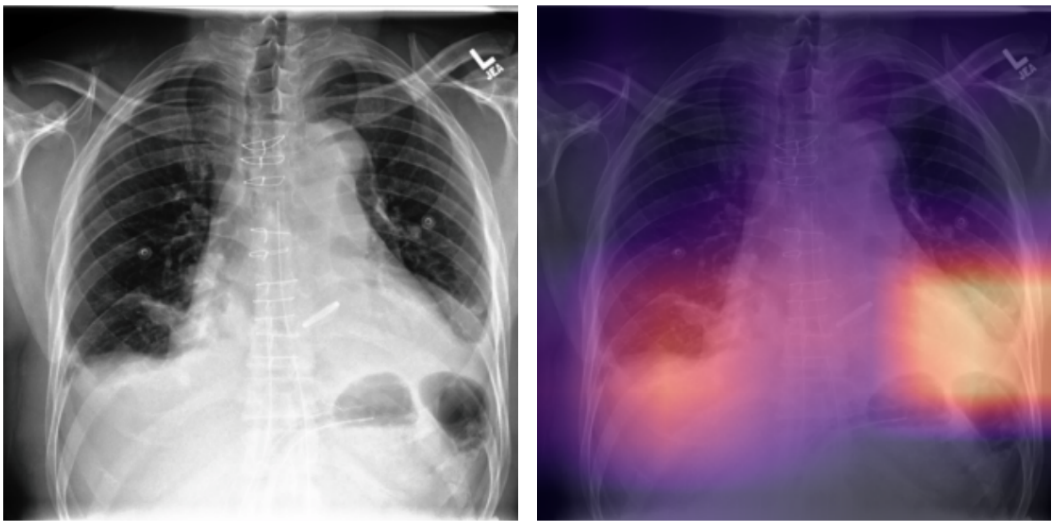


Figure 3.6: Sample images from CheXpert of a patient with Pleural Effusion as taken directly from the data source. Source: [43]

(Table 3.5.5.), Atelectasis (AT), Cardiomegaly (CM), and Pleural Effusion (PE) were selected as labels (see section A.2.1 in appendix 1). These three classes displayed similar number of cases per label and were therefore chosen to establish class balance.

Table 3.5.5: Comparison of labels from CheXpert and NIH Chest X-ray

CheXpert	**NIH Chest X-ray**
Pneumothorax	Enlarged Cardiomediastinum
Edema	**Cardiomegaly**
Pleural Effusion	Consolidation
Pleural Other	Edema
Lesion	Fracture
Atelectacis	Lung Lesion
Pneumonia	Lung Opacity
Consolidation	**Pleural Effusion**
Fracture	Pneumonia
Cardiomegaly	**Atelectasis**
Support device	Pneumothorax
No finding	Pleural Other
	Support Devices
	No Finding

Bounding boxes described in the file are defined as:

$$\text{Bbox}[x, y, w, h]$$

where, $[x, y]$ are the coordinates of the box's top left corner and $[w, h]$ are the width and height of the box respectively.

As outlined in the two-stage approach of the DALUPI algorithm (see Section 3.2.1), the model first maps the relationship between the input x and the selected PI w. In this context, the input x represents chest X-ray images from the two datasets, while the PI consists of bounding boxes. Subsequently, the model integrates information from both the input and the estimated PI to make the final prediction.

3.5.5 Data Preprocessing

Extract and Align Data
Relevant columns are extracted from the NIH Chest X-ray files, including image index, patient ID, view, and domain. Non-overlapping pathologies are filtered out, and the labels are converted into binary labels using a MultiLabelBinarizer from the scikit-learn library. The MultiLabelBinarizer converts the list of labels into binary label indicators. Similarly, for the file containing information on the bounding boxes, the data is aligned, columns are renamed, and non-overlapping pathologies are filtered out. The same steps are carried out for the CheXpert data. The data is aligned by modifying the 'view' columns based on 'Frontal' or 'Lateral' and 'AP/PA' values. Image indices are created after extracting patient IDs. After aligning the data, it is saved onto another CSV file.

CheXpert PI Processing
The CheXpert patient information is loaded from a JSON file. Iterating through each patient's findings and corresponding bounding box annotations, CheXpert patient IDs are mapped to image indices from the aligned dataset. Bounding box coordinates are extracted, and their dimensions are adjusted before recentring the bounding boxes. Visualizations of bounding box annotations are generated as requested. Finally, the processed bounding box annotations are saved into a new CSV file.

Data preparation
For dataset preparation, the source and target datasets were determined based on folder names containing 'NIH' and 'CheXpert'. Datasets were aligned using the CSV alignment function. Subsequently, bounding box annotations for the CheXpert dataset were collected and preprocessed using the aligned indices obtained previously.

3.5.6 Model Adaptations

The experiment was designed for multi-label classification, with four labels: 'No finding', 'Atelectasis', 'Cardiomegaly' and 'Pleural Effusion'. To assess the impact of the bounding boxes on the accuracy of final label predicted, the sizes of the bounding boxes were changed in every iteration of the experiment. Starting with the default size of the bounding boxes as give in the dataset, the sizes were increased and decreased. The Area Under the Curve (AUC) was computed and stored for further analysis. Similar to Experiment I, each experiment was run for 10 epochs

using 5 different hyperparameter settings.

3.6 Evaluation Metrics

The performance of the model can be evaluated using several methods. First, baseline algorithms are implemented on the same input, and the prediction accuracies are recorded. Then, the performance of DALUPI is compared against the performance of the chosen baseline models. In Experiment 1, the metric used is 'accuracy average', while in Experiment 2, 'AUC' is computed.

3.6.1 Accuracy Average

For the "CelebA" experiment, where the task is to classify images of celebrities into male or female categories, the 'accuracy average' metric represents the average accuracy of the model in correctly identifying the gender of the celebrities across all images in the dataset. Accuracy, in the context of binary classification, measures the proportion of correctly classified instances (true positives and true negatives) out of the total instances. In this case, it represents the average rate at which the model predicts the label (male or female) correctly. A high accuracy indicates better performance of the model, meaning it effectively distinguishes between male and female celebrities and assigns their labels correctly. Conversely, a lower accuracy suggests that the model is making more errors in classification.

3.6.2 Receiver Operating Characteristic curve

The ROC curve, or Receiver Operating Characteristic curve, is a graph that illustrates the performance of a classification model. It showcases the model's performance across all classification thresholds and is composed of two parameters: the true positive rate (TPR) and the false positive rate (FPR) [48]. The ROC curve plots TPR against FPR at various classification thresholds. Lowering the threshold results in classifying more items as positive, thereby increasing both true and false positives. Figure 3.7 shows a typical ROC curve.

3.6.3 Area Under the ROC Curve (AUC)

The AUC measures the entire two-dimensional area under the ROC curve. It offers a comprehensive measure of performance across all possible classifications. AUC values range from 0 to 1. A model that makes predictions that are 100% incorrect has an AUC of 0.0, while a model with predictions that are 100% accurate has an AUC of 1.0.

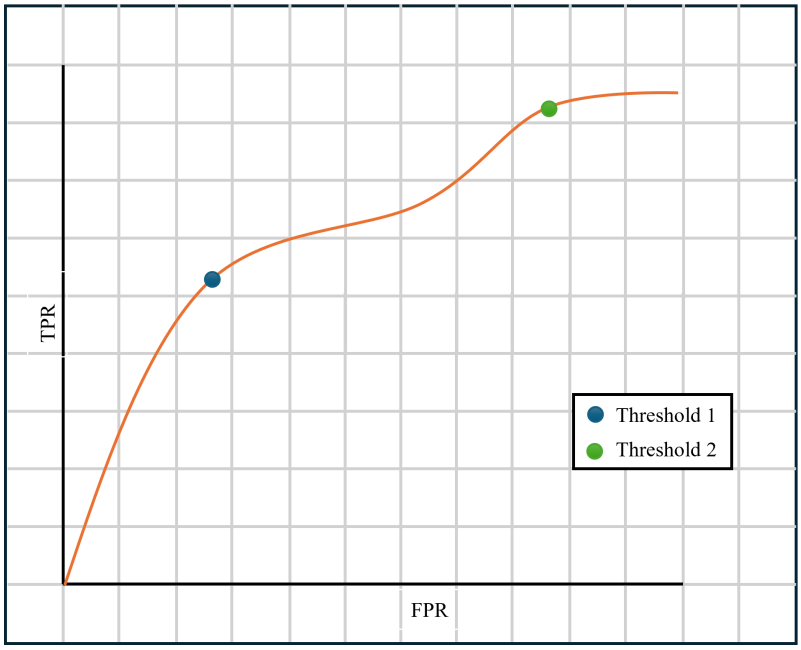


Figure 3.7: Typical ROC curve with TPR vs FPR at different thresholds

4
Results

The results section is structured to address two main categories: the main research questions posed in the thesis and the specific research questions related to each experiment. For the main research questions, the results provide insight into how the methodology implemented contributes to addressing the main objectives of the thesis. Regarding the experiment-specific research questions, the results delve into the outcomes of each experiment in detail. This involves analysing the performance of the model under different conditions, such as varying hyperparameters or input data types, and evaluating how well the model addresses the specific hypotheses outlined for each experiment.

4.1 Binary Attributes as PI

In Experiment I, the analysis focused on utilizing attributes as PI. The experiment involved incrementally increasing the number of attributes per iteration and recording the resulting average accuracies. The outcomes of all 37 iterations were subsequently visualized and analyzed. The following subsection visualises accuracy vs the number of attributes added as PI and further analyses the results.

4.1.1 Accuracy vs Number of Attributes: Visualisation

The experiment involved recording the average accuracy for each iteration. After running all epochs with varying hyperparameters, the accuracies were compared, and the highest accuracy per iteration was retained. The plot of accuracy versus number of attributes illustrates the highest accuracy achieved per iteration, with each iteration adding one attribute to the list of PI inputted to the algorithm. This graph is generated for both the source and target domains to facilitate a comparison and evaluation of the model's behavior. In the source domain, the model incorporates the PI as an input, while in the target domain, the model lacks access to the PI. This comparison of accuracies aims to evaluate the impact of PI on training the model for improved predictions.

4.1.1.1 Attributes in Order of the ACML List

Figure 4.1 shows the scatter plot of accuracy versus the number of attributes, with the attributes ordered according to the ACML list. Examining the average source accuracy, it becomes evident that the inclusion of PI has minimal impact on the model until approximately 24 attributes are incorporated. While some of these 24

attributes exhibit a high correlation with the male label, the later attributes show poor correlation with either label. It is noteworthy that the base accuracy is barely above 50 percent, indicating that the model struggles to learn effectively for a long time. This low base accuracy suggests that the model's initial performance is at chance level, and its learning capacity is limited without the addition of significant features. It might be expected that the model's performance would improve with the inclusion of highly correlated attributes, but the low base accuracy underscores the challenge of learning meaningful patterns from the data without sufficient informative features.

However, the graph illustrates a noticeable absence of significant influence. This behaviour could be attributed to the uneven distribution of male and female images in the source dataset. If the number of male celebrities in the dataset is significantly lower than that of females, attributes correlating with the male label will have minimal influence. As the attributes lean towards higher correlation with the female label, the model demonstrates slightly improved performance, albeit without stabilisation. Moreover, several outliers are noticeable, and there is no discernible trend indicating increasing or decreasing accuracies with the addition of attributes.

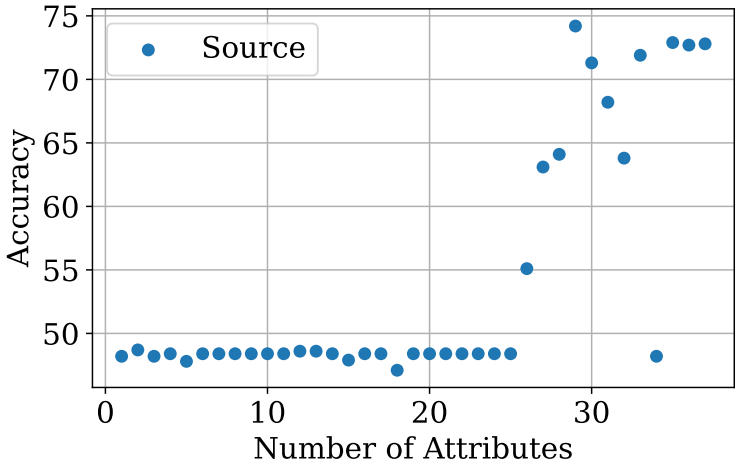


Figure 4.1: Source accuracy plot for DALUPI according to ACML order of attributes

Similar to the average source accuracies, the target accuracies exhibit similar behavior as seen in Figure 4.2. The initial 24 attributes contribute negligibly to the performance of the DALUPI model. The rest of the attributes do not stabilise the performance of the model. This can be interpreted as causing confusion for the model and negatively affecting its performance, resulting in more erroneous predictions rather than PI assisting in improving the average accuracy of this binary classification task.
Overall, the main observation from the average source and target accuracy is that the addition of PI to the model does not enhance its performance. Although there

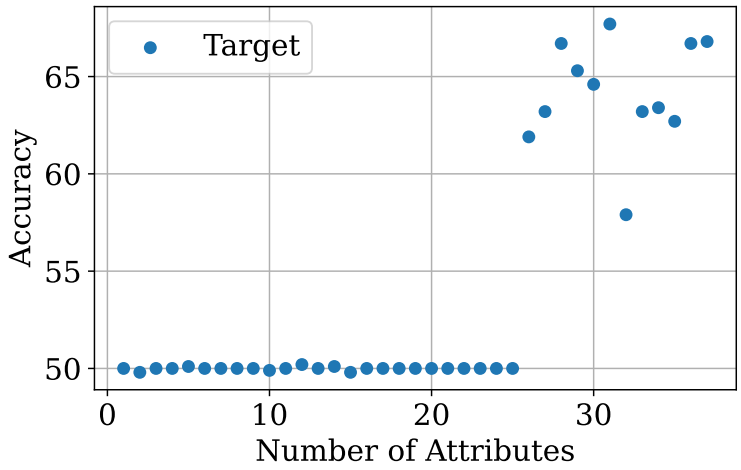


Figure 4.2: Target accuracy plot for DALUPI according to ACML order of attributes

is a sudden increase in prediction accuracy, this improvement is not stable, and the model behaves erratically. The addition of PI in the ACML order does not enhance the performance of the model enough to compare with baseline models and run further experiments on.

4.1.1.2 Attributes in Order of the AOC List

Figure 4.3 shows a scatter plot of the source accuracy when the number of attributes is increased. The analysis of the source graph reveals an initial surge in accuracy, followed by stabilization. However, as the number of attributes increases, the model's accuracy experiences a notable decline. This trend can be attributed to the weak correlation between these additional attributes and the target labels. A lower correlation implies that these attributes contribute less positively to the model's predictions. Consequently, the model becomes increasingly confused, leading to a substantial reduction in accuracy. The presence of attributes with weaker correlations may introduce ambiguity regarding the label, thereby negatively impacting the model's performance.

Moreover, upon closer examination, it becomes evident that the initial three attributes have the most significant impact on the model's performance. The addition of a single attribute has minimal effect, while incorporating the second attribute leads to a substantial improvement in accuracy, approximately by 21%. Furthermore, integrating the top three attributes with the highest overall correlation results in a noticeable performance boost of around 15%. Subsequent iterations of the experiment demonstrate consistent model performance, with the accuracy plateauing at approximately 87% after the inclusion of 26 attributes.

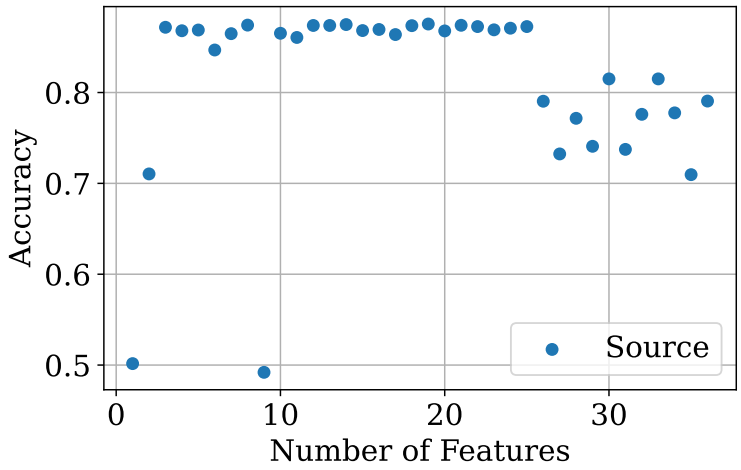


Figure 4.3: Source accuracy plot for DALUPI according to AOC order of attributes

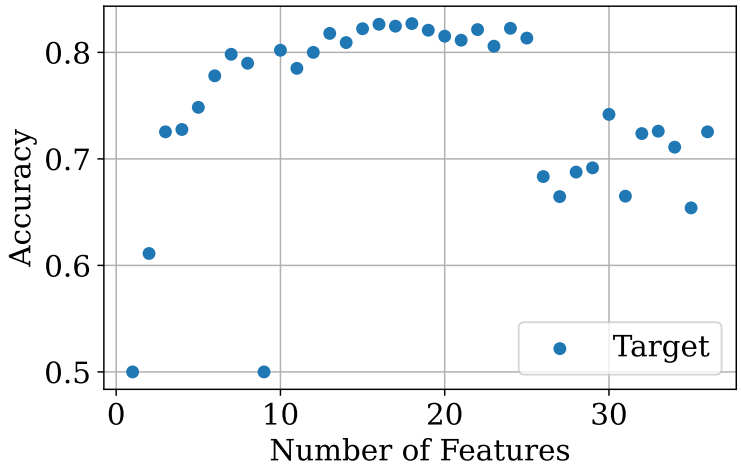


Figure 4.4: Target accuracy plot for DALUPI according to AOC order of attributes

The trend observed in the source domain is further corroborated by the accuracy versus the number of features graph for the target domain as seen in Figure 4.4. Initially, the accuracy rises, and as the number of attributes increases, it achieves a plateau. This plateau indicates the effectiveness of highly correlated attributes. The relationship between accuracy and the number of attributes supports the idea that attributes with strong correlations are crucial for the model's prediction power, whereas attributes with weaker correlations show a reduced impact, as reflected by the plateau in accuracy. The plot also includes an outlier (at attribute number 32), which can be attributed to various factors such as noise in the data or statistical fluctuations arising from the random initialisation of model parameters and stochastic optimisation algorithms. Additionally, the presence of inherent randomness in

the data distribution may contribute to the emergence of outliers in the plot.

4.1.2 DALUPI vs Baseline Algorithms

The performance of DALUPI is then compared to the baseline models described in section 2.2. Performance is also compared with a model designed for binary classification with the addition of and PI. The average target accuracies for each algorithm are listed in table 4.1.1.

Table 4.1.1: Comparison of average accuracy of DALUPI vs other models

Model	Target Accuracy
SL-S	74.4%
DANN	73.9%
MDD	77.3%
DALUPI	82.7%

4.1.2.1 SL-S, DANN and MDD Performance Analysis

The "SL-S" model represents a standard approach in machine learning for binary classification, exclusively using labeled data from the source domain. In contrast, DANN typically requires labeled data from both the source and target domains, or employs UDA techniques. In scenarios with limited labeled data (as in this case) in the target domain, DANN may struggle to adapt effectively, while SL-S can still leverage the available labeled data from the source domain. This difference is reflected in the slight variance in average target accuracies as depicted in the table.

Additionally, the MDD model marginally outperforms SL-S and DANN. This can be attributed to MDD's focus on aligning the distributions of different domains in the feature space, rather than solely relying on labeled samples from the source domain or adversarial domain adaptation techniques. By minimizing the discrepancy between domains, MDD effectively reduces domain shift and increases prediction accuracy, overcoming domain adaptation challenges.

4.1.2.2 DALUPI Performance Analysis

Focusing on DALUPI, it achieves an average target accuracy of 82.7%, outperforming SL-S. While SL-S relies solely on labeled data from the source domain, DALUPI leverages PI, which provides additional insights or context about the data. This enables DALUPI to better align the distributions of the source and target domains, leading to improved adaptation performance.

In contrast to DANN, which aims to learn domain-invariant features through a domain-adversarial training scheme, DALUPI enhances this approach by utilizing PI. By incorporating PI, DALUPI can more effectively exploit the underlying struc-

tures shared between domains, resulting in superior domain adaptation performance compared to DANN.

While MDD primarily focuses on aligning the distributions of different domains in the feature space without the aid of PI, DALUPI leverages PI to gain further insights into the data distribution. By leveraging this additional information, DALUPI can capture more informative features, thereby enhancing its overall performance.

4.1.3 Observations: Experiment I

The DALUPI model demonstrates superior performance compared to other models, achieving a 5.4% increase in prediction accuracy, averaging at 82.7% (Table 4.1.1). This improvement can be attributed to the integration of PI, which effectively addresses domain adaptation challenges. From this experiment, three main observations are drawn:

1. The number of PI significantly influences the model's performance. Insufficient PI may result in minimal impact on the model, while an excessive number can lead to overfitting. This result aligns with hypothesis 1 (3.4.1)
2. The correlation between PI and the label is crucial. Attributes with higher correlations exert a more pronounced effect on the model, while those with lower correlations tend to introduce confusion. This finding is consistent with hypothesis 2 (3.4.1).
3. DALUPI outperforms all baseline models with a significant rise in average accuracy computed in the target domain.

4.2 Bounding Boxes as PI

In Experiment II, the analysis focused on utilizing bounding boxes as PI. The experiment involved adjusting the size of bounding boxes for each iteration of the task. This experiment was designed to evaluate the influence of bounding box size on accurately predicting the label, aiming to gauge the performance of DALUPI in a multi-label classification task.

4.2.1 Estimation of Bounding Boxes

The first stage of the DALUPI algorithm maps a relation between the input (X) and the PI (W). Examples of the bounding boxes estimated are shown in Figure 4.5.
The bounding boxes were computed in various sizes with respect to the original size given as PI and re-centred. Figure 4.5 illustrates the original bounding box in blue, the bounding box reduced to one-quarter of the original size in red, and the bounding box doubled in size in green. Similarly, several boxes were obtained with various sizes.

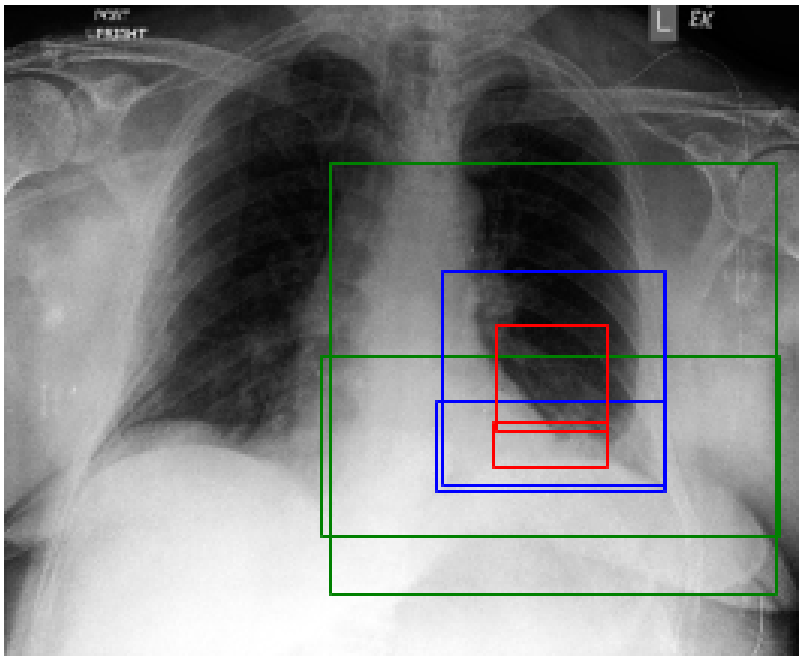


Figure 4.5: Different sizes of bounding boxes (x, 2x and 4x)

4.2.2 AUC of Different Bounding Box Sizes

In tables 4.2.2 and 4.2.3, 'x' represents the size of the bounding box obtained from the dataset. From the table, there is a slight increase in the average AUC of the model. However, it is not significant and could be explained by the presence of noise. Nonetheless, does not indicate poor performance of the model or suggest that the impact of PI is negligible. The model is designed for multilabel classification, and Table 4.2.2 illustrates the AUC obtained for the three labels (label predictions vs ground truth) focused on in this experiment.

Table 4.2.2: Label Specific AUC over Different Bounding Box Sizes in the Source Domain

Size	AUC AT	AUC CM	AUC PE
0.01x	71.56%	80.57%	81.87%
0.25x	71.31%	78.56%	82.92%
0.5x	71.54%	80.19%	78.97%
1x	72.42%	80.61%	79.95%
2x	72.53%	79.20%	83.47%
4x	68.71%	81.62%	80.16%
8x	70.62%	82.79%	80.92%
100x	69.78%	82.25%	79.85%

From table 4.2.2, it can be seen that the AUC values for different pathologies are not consistently maximised at the same bounding box size. For example, AUC-AT peaks at 2x, while AUC-CM and AUC-PE peak at 0.25x. Despite fluctuations in

AUC values across bounding box sizes, the performance remains relatively stable. For instance, the AUC values for AT and PE vary within a range of approximately 1.42% and 0.55%, respectively, across different bounding box sizes. This suggests that the model's performance is robust to changes in bounding box size within this range. However, there are noticeable fluctuations in AUC values across different bounding box sizes, especially for certain labels. For example, the AUC values for the PE label vary from 78.56% to 82.92%, indicating sensitivity to changes in bounding box size.

The same data was collected for the target domain as seen in table 4.2.3. As expected, similar trends were observed in the target domain as well. While there are variations in AUC values among the labels, there is no clear trend indicating the impact of bounding box size on model performance. This suggests that the performance of the model may not be significantly influenced by the size of the bounding box for the target domain.

Table 4.2.3: Label Specific AUC over different Bounding Box Sizes in the Target Domain

Size	AUC AT	AUC CM	AUC PE
0.01x	57.47%	71.57%	58.86%
0.25x	54.33%	70.16%	74.23%
0.5x	55.90%	73.88%	72.57%
1x	54.54%	72.48%	73.15%
2x	55.90%	73.95%	76.55%
4x	54.92%	73.94%	73.12%
8x	55.25%	72.69%	74.11%
100x	53.67%	72.78%	71.34%

Focusing on the two extremes, there are noticeable differences in the AUC values for the bounding box sizes of 0.01x and 100x in the source domain compared to the other sizes. Specifically, the AUC values for the 0.01x size are consistently lower across all label-specific categories, indicating poorer performance in terms of the area under the curve. This suggests that extremely small bounding box sizes may fail to provide sufficient context for the model to accurately classify objects within the images.

Additionally, the variance in results for differently sized bounding boxes between diseases could be influenced by the nature of the diseases themselves. For instance, pleural effusion often presents as a large, diffuse area, making a smaller bounding box less effective for capturing its features. In contrast, conditions like cardiomegaly might have more localized and distinct bounding boxes. The poor performance with the 0.01x size might also be attributed to the inadequacy of such a small box in capturing the relevant features of the diseases, leading to suboptimal model performance. This observation underscores the need to tailor the bounding box sizes to the specific characteristics of each disease for optimal model accuracy.

Conversely, the AUC values for the 100x size also exhibit a decrease in performance compared to other sizes, albeit to a lesser extent. This reduction in performance could be attributed to the excessive inclusion of background information or noise due to the significantly larger bounding box size. Consequently, this may lead to confusion and decreased discriminative ability in the model's predictions.

The table 4.2.4 presents a comparison of the average accuracy of DALUPI against other models across different datasets. DALUPI outperforms all other models, achieving the highest target accuracy of 71.75%. Specifically, in the AT dataset, DALUPI's accuracy is 55%, compared to 57% for SL-T, 55% for SL-S, 53% for DANN, and 49% for MDD. In the CM dataset, DALUPI significantly surpasses the other models with an accuracy of 72%, while SL-T, SL-S, DANN, and MDD have accuracies of 59%, 61%, 55%, and 51%, respectively. For the PE dataset, DALUPI also leads with an accuracy of 74%, whereas the competing models have accuracies of 79% for SL-T, 73% for SL-S, 55% for DANN, and 51% for MDD.

Table 4.2.4: Comparison of average accuracy of DALUPI vs other models

Model	**AT**	**CM**	**PE**
SL-T	57%	59%	79%
SL-S	55%	61%	73%
DANN	53%	55%	55%
MDD	49%	51%	51%
DALUPI	55%	72%	74%

The experimental results diverged from the hypothesis that models trained with 0.01 and 100 times the size of the original bounding box would perform similarly to a model trained without any PI such as the SL-S model. Contrary to expectations, these models exhibited higher accuracies across all metrics compared to SL-S. One possible explanation for this discrepancy could be attributed to the differences in model architecture. While SL-S relies on ResNet, DALUPI employs faster R-CNN. This architectural disparity likely contributed to the observed variations in accuracy, as Faster R-CNN may inherently provide better representations and feature extraction capabilities compared to ResNet.

While the 0.01x bounding box does not provide specific spatial information to the model, it indicates where to look for the label within the entire image. Conversely, the 100x bounding box fails to provide information about specific spatial markers or the location of potential label areas. This difference significantly contributes to the superior performance of the 0.01x bounding box model compared to the 100x bounding box and the SL-S model.

To investigate the impact of the placement of the bounding box (3.5.1), the experiment was conducted similar to experiment two, but the location of the bounding box was shifted slightly to the left, than the original location predicted.

Table 4.2.5: Source accuracy when bounding box locations are moved

Location of BB	AT	CM	PE
Original	72.42%	80.61%	79.95%
2 pixels to the left	70.03%	81.76%	79.32%
5 pixels to the left	72.53%	81.58%	79.3%

As indicated by Table 4.2.5, the accuracy of the model demonstrates varying trends when the bounding box positions are altered. Specifically, moving the bounding box 2 pixels to the left improves accuracy for the CM dataset to 81.76%, slightly higher than the original position at 80.61%. Additionally, the best accuracy for the AT dataset is observed when the bounding box is shifted 5 pixels to the left, reaching 72.53%. Conversely, the accuracy for the PE dataset remains largely unaffected by the shift. This is likely due to the inherent characteristics of the diseases: PE typically presents as a more diffuse condition, making its boundaries less sensitive to minor shifts. In contrast, CM often involves more localized and distinct features, allowing for slight adjustments in bounding box positions to enhance detection accuracy. The AT dataset's performance improvement with a 5-pixel shift suggests that the model benefits from slight adjustments that better capture the disease's focal points. This underscores the importance of considering disease-specific characteristics when fine-tuning model parameters.

Table 4.2.6: Target accuracy when bounding box locations are moved

Location of BB	AT	CM	PE
Original	54.54%	72.48%	73.15%
2 pixels to the left	54.24%	71.2%	69.24%
5 pixels to the left	54.16%	70.8%	69.3%

5
Discussions

Based on the experiments conducted and the results obtained, the study provides valuable insights into the role of PI in domain adaptation for image classification. This chapter delves into discussions about the results and the experiment specific research questions, as well as the overall research questions investigated in this thesis.

5.1 Experiment-Specific Inferences

The first experiment demonstrated a significant impact of incorporating PI into the DALUPI model. Analysis indicated that integrating PI led to increased accuracy and stabilized prediction accuracies. Notably, the sequence of attribute integration influenced the number of attributes required for stable output accuracy. These results align with the hypotheses proposed, emphasizing the importance of both the quantity and correlation of PI in enhancing model performance.

Previous studies have also highlighted the importance of PI in improving model performance in domain adaptation tasks. The findings of this experiment align with existing literature, emphasizing the significance of considering both the amount and correlation of PI with output labels for effective domain adaptation.

The second experiment focused on utilizing bounding boxes as PI, with variations in their dimensions across each iteration. Analysis of different bounding box sizes revealed variability in model performance, indicating that bounding box size does indeed impact accuracy, albeit to a minimal extent. Notably, employing the faster R-CNN for predicting bounding boxes and subsequently labeling multi-label chest X-ray experiments outperformed the conventional ResNet model. This contradicts previous literature, which typically favors the standard ResNet model. Additionally, the findings underscore the significance of bounding box positioning in enhancing model performance and reducing the number of false negatives. Even minor alterations in predicted bounding box placement can affect model accuracy.

Despite variations in performance across different bounding box sizes, DALUPI consistently outperformed baseline models, underscoring its effectiveness in utilizing PI for domain adaptation. While previous research has explored the use of bounding boxes as PI in various contexts, few studies have specifically investigated their impact on domain adaptation in medical image classification. This experiment contributes novel insights into the role of bounding box size as PI in domain adaptation,

providing valuable information for researchers and practitioners in the field.

The experimental results contribute significantly to theoretical understanding by validating the hypotheses proposed and emphasizing the crucial role of PI in domain adaptation tasks. The findings affirm that both the quantity of PI provided and its correlation with output labels are key determinants of model performance, thereby supporting the fundamental principles of domain adaptation. This validation enriches existing theories of domain adaptation and machine learning by furnishing empirical evidence of the pivotal role played by PI in enhancing model adaptability across diverse domains. Moreover, the study underscores the importance of strategically considering PI attributes during model training to achieve stable output accuracy and effectively mitigate domain shift.

5.2 Overall Findings and Implications

The thesis explored three main questions related to the application of PI in machine learning algorithms to address domain adaptation issues in imaging fields. The findings demonstrate that the DALUPI model, which utilizes PI, surpasses other models, thus underscoring the beneficial effects of supplementary information on the model. The initial experiment reveals that both the quantity and the quality of PI significantly influence the model's effectiveness. In this context, 'quality' denotes the degree of correlation between the PI and the target labels that the model predicts. A higher correlation means less PI is needed for better predictions and improved accuracy of the model. On the other hand, PI with lower correlation necessitates the introduction of more PI, which tends to confuse rather than enhance the model's learning and accuracy.

The second research question explored the extent to which the accuracy of PI, gauged by factors like detail level, affects the efficacy of image classifiers in environments with significant distribution shifts. The second experiment aimed to address this issue. Different bounding box sizes either decrease or increase the detail level provided as PI. Smaller boxes indicate the location of a potential label (disease/ailment) but offer minimal area for information extraction. However, the model identifies relevant areas to focus on. In contrast, larger bounding boxes deliver extensive information regarding the area to search for feature markers. Yet, they might include unnecessary details that overwhelm the model. Although this experiment did not fully support the initial hypothesis, it demonstrated that neither too small nor too large bounding boxes are ideal for making precise predictions. Moreover, the level of detail provided by a single type of PI can vary, including essential feature markers, areas to concentrate on, and areas to ignore.

The third research question aimed to assess the algorithm's robustness and its potential applicability beyond medical imaging domains. To this end, the DALUPI algorithm was evaluated using two distinct types of PI on disparate datasets. Specifically, binary attributes served as PI for the CelebA dataset, while bounding boxes

functioned as PI for chest X-ray data sourced from two separate datasets. Remarkably, DALUPI consistently outperformed all other tested models in both scenarios. These results underscore the algorithm's robustness and its capacity for broader application beyond medical imaging domains. They suggest the feasibility of extending the algorithm to accommodate diverse data types beyond images, highlighting its versatility and potential for diverse applications.

5.3 Limitations and Future Scope

This study is constrained by certain limitations, which expand the scope for future research. These limitations are broadly divided into 5 areas.

5.3.1 Generalization to Other Domains

In this thesis, the focus is primarily on medical image classification, with an extension to the binary classification of male and female celebrities. However, this specific application may not generalize to other domains without further validation. Future research could explore the applicability of DALUPI and PI in diverse domains such as natural language processing, computer vision, or financial analysis.

5.3.2 Scalability

The ability to scale the suggested approaches to bigger datasets or practical applications warrants further exploration. Expanding to larger datasets could introduce computational difficulties or necessitate enhancements in the algorithm's implementation.

5.3.3 Noise and Variability

Evaluating the resilience of DALUPI against noisy or inconsistent data is crucial. Datasets in real-world conditions frequently exhibit noise, gaps, or irregularities that could impact the algorithm's effectiveness. Conducting a robustness analysis will help determine the dependability and steadiness of DALUPI under practical conditions.

5.3.4 Scope of PI

While this study examines the impact of two different types of PI, there may be other types of PI that could further enhance model performance. Exploring additional forms of PI could provide deeper insights into its effectiveness.

5.3.5 Assumption of Domain Shift

The effectiveness of domain adaptation methods, including DALUPI, is based on the assumption of domain shift between the source and target domains. However, this assumption may not universally apply to all scenarios, leading to potential

limitations in model performance in certain contexts. For instance, in situations where the source and target domains are highly similar, such as diagnosing a disease in patients from the same demographic and geographical background, the need for extensive domain adaptation might be minimal. Conversely, in cases involving stark differences in data distributions, such as imaging data from different medical centers with varied equipment and protocols, the domain shift assumption becomes critical. For example, a model trained on X-ray images from one hospital may struggle with images from another hospital if the imaging protocols or patient populations differ significantly. This highlights the need for careful consideration of domain characteristics and the potential limitations of domain adaptation approaches in diverse real-world applications.

5.4 Societal, Ethical and Ecological considerations

The proposed project emphasizes the significance of ethical and environmental concerns. Conducting the research in a manner that respects the broader implications and responsibilities associated with this work is crucial to its success. The following subsections identify and address some key questions related to ethics and environmental impact.

5.4.1 Societal Impact

1. *Benefiting Society*
   This comprehensive study on the robustness of the DALUPI approach to mitigate challenges of domain adaptation will help improve diagnostic accuracy in various domains, especially healthcare, which can lead to more effective treatments and better patient outcomes.

2. *Addressing Societal Concerns*
   We acknowledge concerns regarding privacy and bias; To address them, we plan on utilising anonymous and protect sensitive data and use de-identifiable datasets available. We also intend to implement fairness-sensitive algorithms to prevent any kind of bias.

5.4.2 Ethical Considerations

1. *Ethical Guidelines and Principles*
   We are dedicated to abiding by accepted moral standards for managing data. This dedication extends to all aspects of our study, including data collection, analysis, and reporting. Integrity, respect for privacy, and appropriate data handling are the foundations of our approach.

2. *Informed Consent, Data Privacy and Anonymizatio*n
   The datasets used for this study are obtained after undertaking special certifications in data handling and ethical consideration.

After obtaining informed consent and rigorous data anonymization, the datasets used for this study are made available to the public by the respective data owner. The datasets used for this study are curated by the respective body after obtaining informed consent and only after undergoing rigorous data anonymization.

5.4.3 Ecological Impact

Environmental Responsibility

We are dedicated to reducing our ecological impact by utilising energy-efficient hardware, cloud computing strategies, and effective data storage. We want to reduce the effect of our study on the environment by using eco-friendly methods.

In this thesis, ethical guidelines were strictly adhered to, substantial contributions were made to the discipline, and societal interests and values were maintained by actively engaging with the sociological, ethical, and environmental dimensions involved.

6
Conclusion

This thesis is centered on investigating the role of PI in addressing domain adaptation challenges, focusing on implementing the DALUPI algorithm across diverse datasets featuring various types of PI. Two experiments were conducted to address three key research questions. Based on the results presented in Section 4 and the discussions outlined in Section 5, several important conclusions can be drawn.

Robustness of the model

The implemented model exhibits robustness across diverse domains and data types. While primarily focusing on image data, the extension of the DALUPI algorithm to two distinct image datasets—celebrity faces and medical chest X-ray images—illustrates its versatility. Notably, despite data sourced from various hospitals, the DALUPI model consistently outperformed models designed for domain adaptation challenges, indicating its potential for broader applications beyond medical imaging.

Number and relevance of PI matters

The significance of PI is apparent in impacting both quantity and quality, crucial for model performance. Initially, an increase in PI positively affects accuracy, but beyond a certain point, the model stabilizes or overfits to accommodate excessive PI. Moreover, the correlation between PI and target labels emerges as vital, with aligned PI requiring less volume for accurate predictions. Conversely, irrelevant PI confuses the model, potentially deteriorating performance. Achieving a balance between the number and relevance of PI is essential for optimizing model performance.

Accuracy of PI

The experiments underscored the significance of the accuracy of PI in influencing model performance. While bounding box size did not notably affect model performance, precise placement emerged as crucial for accurate diagnosis. Misalignment between bounding boxes and relevant features could result in false negatives, emphasizing the importance of detailed PI perception by the model.

The practical implications of these findings are profound, particularly in medical image classification, where domain adaptation faces challenges of data scarcity and distribution shifts. Incorporating PI into machine learning models offers promising avenues for improving real-world applications in medical diagnosis and treatment

planning. By harnessing additional contextual information provided by PI, models can better generalize their learning to new domains, leading to more precise and reliable predictions, ultimately enhancing patient care.

As this thesis concludes, it highlights the pivotal role of PI in enhancing machine learning algorithms' adaptability to diverse domains. The DALUPI model's efficacy in harnessing PI to elevate model performance underscores the critical importance of meticulously considering the quantity, quality, and relevance of PI. Looking ahead, the insights gleaned from this research offer valuable guidance for advancing machine learning methodologies, fostering more dependable and efficient solutions for intricate domain adaptation tasks.

Bibliography

[1] American Journal of Roentgenology. (n.d.). Automated Detection of COVID-19 Cases Using Deep Neural Networks with X-ray Images. Retrieved from https://ajronline.org/doi/full/10.2214/AJR.12.10375

[2] Healthcare Artificial Intelligence Market to Register Commendable 40 Percent CAGR Over 2017-2024. (2018). Retrieved from https://www.mpo-mag.com/contents/view_online-exclusives/2018-04-16/healthcare-artificial-intelligence-market-to-register-commendable-40-percent-c Accessed: September 2023.

[3] Healthcare artificial intelligence market to record a CAGR of 40 percent over 2017-2024. (n.d.). Retrieved from https://www.healthcarefacilitiestoday.com/posts/Healthcare-artificial-intelligence-market-to-record-a-CAGR-of-40-percent-over Accessed: September 2023.

[4] Vladimir Vapnik and A. Vashist. (2009). A new learning paradigm: Learning using privileged information. *Neural Networks*, **22**(5), 544–557. https://doi.org/10.1016/j.neunet.2009.06.042

[5] Qinmu Dou, Cheng Ouyang, Yuankai Chen, Hao Chen, Pheng-Ann Heng. (2018). Deep Domain Adaptation for Prostate MRI Segmentation. *Medical Image Analysis*, **49**, 69-80.

[6] Muhammad Imran Razzak, Saeed Anwar Naz, Ahmad Zaib. (2018). Domain Adaptation for Histopathology Image Analysis: A Comprehensive Review. *IEEE Reviews in Biomedical Engineering*, **11**, 206-217.

[7] Majdi Ben Salah, Mohamed Ammar Boudjelal, Mohamed Hedi Bedoui. (2019). Domain Adaptation for Brain Lesion Segmentation: A Solution to Handle Data Heterogeneity. In *Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence*, 417-427.

[8] Xiaopeng Zheng, Jianming Shi, Shuai Ying, Qinghui Zhang, Yun Li. (2016). Improving Single-Modal Neuroimaging Based Diagnosis of Brain Disorders via Boosted Privileged Information Learning Framework. In *Machine Learning in Medical Imaging*, 98-106.

[9] Xiaopeng Zheng, Jianming Shi, Qinghui Zhang, Shuai Ying, Yun Li. (2017). Improving MRI-based diagnosis of Alzheimer's disease via an ensemble privileged information learning algorithm. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 456-459. https://doi.org/10.1109/ISBI.2017.7950559

[10] Zhi Gao, Shun Miao Wu, Zhennan Liu, Junjie Luo, Hong Zhang, Mingliang Gong, Shuyu Li. (2019). Learning the implicit strain reconstruction in ultrasound elastography using privileged information. *Medical Image Analysis*, **58**, doi: 10.1016/j.media.2019.101534

[11] Fanfan Ye, Jianan Pu, Jingyi Wang, Yang Li, Hongzhi Zhang, Hongyuan Zha. (2017). Glioma grading based on 3D multimodal convolutional neural network and privileged learning. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 759-763. doi: 10.1109/BIBM.2017.8217751

[12] Anton Breitholtz, Alexander Matsson, Fredrik D. Johansson. (2023). Unsupervised domain adaptation by learning using privileged information. *arXiv preprint arXiv:2303.09350*.

[13] Hidetoshi Shimodaira. (2000). Improving predictive inference under covariate shift by weighting the log likelihood function. *Journal of Statistical Planning and Inference*, **90**(2), 227–244.

[14] Alex D'Amour, Purnamrita Sarkar, Victor Veitch. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, **221**(2), 644-654.

[15] Fredrik D. Johansson, David Sontag, Rajesh Ranganath. (2019). Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR.

[16] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Ciurea-Ilcus, Silviana, Chute, Chris, Marklund, Henrik, Haghgoo, Bijan, Ball, Robyn, Shpanskaya, Katie, Seekins, Jay, Mong, David A., Halabi, Safwan S., Sandberg, Johan K., Jones, Ricky, Larson, David B., Langlotz, Curtis P., Patel, Bhavik N., Lungren, Matthew P., Andrew Y. Ng. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *arXiv preprint arXiv:1901.07031*.

[17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald M. Summers. (2017). ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Alistair E. W. Johnson, Tom J. Pollard, Roger G. Mark, Seth J. Berkowitz, Steven Horng. (2019). MIMIC-CXR Database (version 2.0.0). PhysioNet. doi: 10.13026/C2JT1Q.

[19] Alistair E.W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Steven J. Horng. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, **6**, 317. doi: 10.1038/s41597-019-0322-0

[20] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, H. Eugene Stanley. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, **101**(23), e215–e220. https://doi.org/10.13026/C2JT1Q

[21] IXI Dataset. (n.d.). Brain Development. Retrieved from http://brain-development.org/ixi-dataset/

[22] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, et al. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, **34**(10), 1993-2024. doi: 10.1109/TMI.2014.2377694

[23] Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017). Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Nature Scientific Data*, **4**, 170117. https://doi.org/10.1038/sdata.2017.117

[24] Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *arXiv preprint arXiv:1811.02629*.

[25] PhysioNet. (n.d.). MIMIC II Clinical Overview. Retrieved from https://archive.physionet.org/mimic2/mimic2_clinical_overview.shtml Accessed: September 11, 2023.

[26] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, **17**(59), 1–35. http://jmlr.org/papers/v17/15-239.html

[27] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., Smola, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research*, **13**, 723-773.

[28] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, **79**(1-2), 151-175.

[29] Long, M., Cao, Y., Cao, Z., Wang, J., Jordan, M. I. (2019). Transferable Representation Learning with Deep Adaptation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**(12), 3071-3085. https://doi.org/10.1109/TPAMI.2018.2868685

[30] Pan, S. J., Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, **22**(10), 1345-1359. https://doi.org/10.1109/TKDE.2009.191

[31] Saleh, M. A., Ali, A. A., Ahmed, K., Sarhan, A. M. (2022). A Brief Analysis of Multimodal Medical Image Fusion Techniques. *Electronics*, **12**(1), 97. https://doi.org/10.3390/electronics12010097

[32] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V. (2016). Domain-Adversarial Training of Neural Networks. arXiv preprint arXiv:1505.07818. Retrieved from https://arxiv.org/abs/1505.07818

[33] Rajkumar, Sumit. (2021). Domain Adversarial Training of Neural Networks: 2016 One-Minute Summary.

[34] IBM. (n.d.). Image Segmentation. Retrieved from https://www.ibm.com/topics/image-segmentation#:~:text=Image%20segmentation%20is%20a%20computer,faster%2C%20more%20advanced%20image%20processing.

[35] Hewlett Packard Enterprise. (n.d.). Convolutional Neural Network (CNN). Retrieved from https://www.hpe.com/se/en/what-is/convolutional-neural-network.html#:~:text=CNNs%20are%20used%20for%20image,shapes%2C%20and%20objects%20within%20images.

[36] ScienceDirect. (n.d.). Medical image analysis: Methods. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/B9780323265683000051

[37] Scholarpedia. (n.d.). Functional Imaging. Retrieved from http://www.scholarpedia.org/article/Functional_imaging#:text=Functional%20imaging%20is%20the%20study,(PET)%20or%20Optical%20Imaging.

[38] Monocosmo. (2024). How Maximum Mean Discrepancy Is Utilised Efficiently - Part 1: Machine Learning.

[39] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[40] Python.org. (n.d.). Retrieved from https://www.python.org. Accessed on: December 20, 2023.

[41] Swedish National Infrastructure for Computing (SNIC). (n.d.). Alvis - SNIC Science Cloud. Retrieved from https://www.snic.se/resources/compute-resources/alvis/index.html.

[42] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R. M. (2017). ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on

Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.

[43] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R. L., Shpanskaya, K. S., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., Ng, A. Y. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *CoRR*, abs/1901.07031. Retrieved from `http://arxiv.org/abs/1901.07031`

[44] Strickland, N. H. (2000). PACS (picture archiving and communication systems): filmless radiology. *Arch Dis Child*, 83(1), 82-6. doi: 10.1136/adc.83.1.82. PMID: 10869010; PMCID: PMC1718393.

[45] Jones, J., Weerakkody, Y., Worsley, C., et al. (Year). Atelectasis (summary). Reference article, Radiopaedia.org. Retrieved from `https://radiopaedia.org/articles/513`. DOI: `https://doi.org/10.53347/rID-51373`

[46] Mayo Clinic. (n.d.). Atelectasis - Symptoms and causes. Retrieved from `https://www.mayoclinic.org/diseases-conditions/atelectasis/symptoms-causes/syc-20369684`.

[47] LearningRadiology.com. (n.d.). Right Lower Lobe Atelectasis. Retrieved from `https://learningradiology.com/notes/chestnotes/rllatelectasis.htm`.

[48] Google Developers. (n.d.). Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC). Retrieved from `https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc`.

[49] Code Institute. (n.d.). What Are Libraries in Python? Retrieved from `https://codeinstitute.net/se/blog/what-are-libraries-in-python/#:~:text=Using%20a%20Python%20library%20is,Libraries%20are%20collections%20of%20modules`

[50] MMLAB, The Chinese University of Hong Kong. (n.d.). CelebA Dataset. Retrieved from `https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`.

[51] Google Cloud. (n.d.). What is Supervised Learning? Retrieved from `https://cloud.google.com/discover/what-is-supervised-learning#:~:text=Supervised%20learning%20is%20a%20category,the%20input%20and%20the%20outputs`.

[52] Google Cloud. (n.d.). What is Unsupervised Learning? Retrieved from `https://cloud.google.com/discover/what-is-unsupervised-learning`.

[53] IBM. (n.d.). Semi-Supervised Learning. Retrieved from `https://www.ibm.com/topics/semi-supervised-learning#:~:text=Semi-supervised%20learning%20is%20a,for%20classification%20and%20regression%20tasks`.

[54] Analytics Vidhya. (2019, August). A Comprehensive Guide to K-Means Clustering. Retrieved from https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/.

[55] Medium. (n.d.). ML Part 5 - Clustering. Retrieved from https://medium.com/@avicsebooks/ml-part-5-clustering-00d30a977b50.

[56] Amazon Web Services. (n.d.). What is Data Labeling? Retrieved from https://aws.amazon.com/what-is/data-labeling/#:~:text=In%20machine%20learning%2C%20a%20properly,highly%20accurate%20data%20labeling%20is.

[57] Graphite Note. (n.d.). Data Labeling for Machine Learning Success. Retrieved from https://graphite-note.com/data-labeling-for-machine-learning-success/.

[58] BuiltIn. (n.d.). Regression in Machine Learning. Retrieved from https://builtin.com/data-science/regression-machine-learning.

[59] DataCamp. (n.d.). Classification in Machine Learning. Retrieved from https://www.datacamp.com/blog/classification-machine-learning.

[60] AltexSoft. (n.d.). Semi-Supervised Learning: Types, Techniques, and Applications. Retrieved from https://www.altexsoft.com/blog/semi-supervised-learning/.

[61] Symbl AI. (n.d.). Understanding Semi-Supervised Learning. Retrieved from https://symbl.ai/developers/blog/understanding-semi-supervised-learning/.

[62] Google Cloud. (n.d.). Deep Learning vs Machine Learning. Retrieved from https://cloud.google.com/discover/deep-learning-vs-machine-learning#:~:text=Essentially%2C%20deep%20learning%20can%20learn%2Crequires%20significantly%20more%20computational%20power..

[63] GeeksforGeeks. (n.d.). Convolutional Neural Network (CNN) in Machine Learning. Retrieved from https://www.geeksforgeeks.org/convolutional-neural-network-cnn-in-machine-learning/.

[64] Towards Data Science. (n.d.). Convolutional Neural Networks Explained. Retrieved from https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939.

[65] UpGrad. (n.d.). Basic CNN Architecture. Retrieved from https://www.upgrad.com/blog/basic-cnn-architecture/.

[66] Alzubaidi, L., Zhang, J., Humaidi, A. J., et al. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1), 53. doi: [10.1186/s40537-021-00444-8](https://doi.org/10.1186/s40537-021-00444-8)

[67] R. Girshick, J. Donahue, T. Darrell, & J. Malik. (2013). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. doi: 10.1109/CVPR.2014.81

[68] Zhang, Aston, Zachary C. Lipton, Mu Li, & Alexander J. Smola. (2022). *d2l.ai: Dive into Deep Learning*. Retrieved April 28, 2024, from `https://d2l.ai/`

[69] GeeksforGeeks. (n.d.). VGG-16 CNN Model. Retrieved from `https://www.geeksforgeeks.org/vgg-16-cnn-model/`

[70] Keras Documentation. (n.d.). VGG. Retrieved from `https://keras.io/api/applications/vgg/`

[71] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.

[72] BuiltIn. (n.d.). Backpropagation in Neural Networks: The Math Behind the Curtain. Retrieved from `https://builtin.com/machine-learning/backpropagation-neural-network`

[73] Towards Data Science. (n.d.). Loss Functions and Their Use in Neural Networks. Retrieved from `https://towardsdatascience.com/loss-functions-and-their-use-in-neural-networks-a470e703f1e9`

[74] BuiltIn. (n.d.). Gradient Descent in Machine Learning: What Is It and How Does It Work? Retrieved from `https://builtin.com/data-science/gradient-descent`

[75] Towards Data Science. (n.d.). Understanding the Mathematics Behind Gradient Descent. Retrieved from `https://towardsdatascience.com/understanding-the-mathematics-behind-gradient-descent-dde5dc9be06e`

[76] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. Retrieved from `https://arxiv.org/pdf/1512.03385`

[77] Ibteda Azeem. (n.d.). *Understanding ResNet Architecture: A Deep Dive into Residual Neural Network*.

[78] DataGen. (n.d.). ResNet: An in-depth guide. Retrieved from `https://datagen.tech/guides/computer-vision/resnet/`

[79] Papers with Code. (n.d.). ResNet. Retrieved from `https://paperswithcode.com/method/resnet`

[80] Samar Bashath, Nadeesha Perera, Shailesh Tripathi, Kalifa Manjang, Matthias Dehmer, & Frank Emmert-Streib. (2021). A data-centric review of deep transfer learning with applications to text data. *Information Sciences*, 585. doi: 10.1016/j.ins.2021.11.061

[81] Hong Zeng, Xiufeng Li, Gianluca Borghini, Yue Zhao, Pietro Aricò, Gianluca Di Flumeri, Nicolina Sciaraffa, Wael Zakaria, Wanzeng Kong, & Fabio Babiloni. (2021). An EEG-Based Transfer Learning Method for Cross-Subject Fatigue Mental State Prediction. *Sensors*, 21(3), 2369. doi: 10.3390/s21072369

[82] Hao Guan & Mingxia Liu. (2021, October 4). Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Transactions on Biomedical Engineering*. doi: 10.1109/TBME.2021.3117407

[83] Dapeng Hu, Mi Luo, Jian Liang, & Chuan-Sheng Foo. (2024). Simplifying and Stabilizing Model Selection in Unsupervised Domain Adaptation. Retrieved from `https://openreview.net/forum?id=S6Xf70Y5CJ`

[84] Sk Miraj Ahmed, Dripta S. Raychaudhuri, Samet Oymak, & Amit K. Roy-Chowdhury. (2022). Chapter 5 - Source distribution weighted multisource domain adaptation without access to source data. Retrieved from `https://doi.org/10.1016/bs.host.2022.12.001`

[85] Laila El Jiani, Sanaa El Filali, & El Habib Benlahmer. (2022). Overcome medical image data scarcity by data augmentation techniques: A review. In *2022 International Conference on Microelectronics (ICM)* (pp. 21-24). doi: 10.1109/ICM56065.2022.10005544

[86] Dimitrios Theodoropoulos. The Challenge of Lack of Medical Datasets in Deep Learning for Medical Imaging.

[87] Xiaofeng Liu, Chaehwa Yoo, Fangxu Xing, Hyejin Oh, Georges El Fakhri, Je-Won Kang, & Jonghye Woo. (2022). Deep Unsupervised Domain Adaptation: A Review of Recent Advances and Perspectives. arXiv preprint arXiv:2208.07422.

[88] IBM. (n.d.). Supervised vs. Unsupervised Learning. Retrieved from `https://www.ibm.com/blog/supervised-vs-unsupervised-learning/`

[89] Towards Data Science. Detecting Covariate Shift: A Guide to the Multivariate Approach.

[90] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)", *IEEE Transactions on Medical Imaging*, 34(10), 1993-2024 (2015) DOI: 10.1109/TMI.2014.2377694

[91] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features", *Nature Scientific Data*, 4:170117 (2017) DOI: 10.1038/sdata.2017.117

[92] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, et al., "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge", *arXiv preprint arXiv:1811.02629* (2018)

[93] Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, Mark RG, Horng S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*. 2019;6.

[94] Johnson AE, Pollard TJ, Greenbaum NR, Lungren MP, Deng C-Y, Peng Y, Lu Z, Mark RG, Berkowitz SJ, Horng S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*. 2019.

[95] Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). *PhysioNet*. 10.13026/C2XW26

[96] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035.

[97] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation [Online]*, 101(23), pp. e215–e220.

[98] IXI Dataset. Retrieved from `https://brain-development.org/ixi-dataset/`

[99] MIMIC-II Clinical Overview. Retrieved from `https://archive.physionet.org/mimic2/mimic2_clinical_overview.shtml`

[100] Cleveland Clinic. (n.d.). Pleural Effusion. Retrieved April 28, 2024, from `https://my.clevelandclinic.org/health/diseases/17373-pleural-effusion`.

[101] National Center for Biotechnology Information. (n.d.). Cardiomegaly. In *StatPearls*. Retrieved April 28, 2024, from `https://www.ncbi.nlm.nih.gov/books/NBK542296/#:~:text=Cardiomegaly%20is%20an%20umbrella%20designation,means%20enlargement%20of%20the%20heart`.

[102] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. *arXiv preprint arXiv:1506.02640*, 2016. `http://arxiv.org/abs/1506.02640`.

[103] GeeksforGeeks. Residual Networks (ResNet) in Deep Learning. *GeeksforGeeks*, 2023. `https://www.geeksforgeeks.org/residual-networks-resnet-deep-learning/`.

A

Appendix 1

A.1 Types of ML

Generally, the training of any AI model necessitates the use of labeled data for comparison to assess the model's precision. Labeled data is essential in the training of AI models in various sectors, acting as the cornerstone for supervised machine learning algorithms. These algorithms depend on it to learn the connections between input variables and their corresponding outputs. In the absence of labeled data, these algorithms would miss crucial reference points needed for making precise predictions or classifications. Labeling data typically requires human annotation or expert input, assigning each data point a specific category or result [56] [57]. This labeled dataset then acts as a standard for measuring the effectiveness of the model and ensuring its accuracy and dependability. Based on the availability of labelled data, ML algorithms can be categorised in three ways.

A.1.1 Supervised Machine Learning

Supervised machine learning employs labelled datasets to train algorithms to predict outcomes and identify patterns [51]. The data includes examples of both inputs (features) and accurate outputs (labels). These algorithms examine a vast collection of these training pairs to deduce the desired output value when making predictions on unfamiliar data. Supervised machine learning can be broadly divided into two categories based on the overall goal of the algorithm.

A.1.1.1 Regression

Regression techniques are applied in scenarios where the output variable is of a continuous or numerical nature. Instead of predicting a class label, these methods forecast a continuous quantity. Regression models ascertain the correlation between input variables and the continuous response variable, facilitating the prediction of numerical results. Common examples of regression techniques are linear regression, polynomial regression, decision trees, and neural networks [58].

A.1.1.2 Classification

Classification algorithms aim to predict the category or class label of a new observation based on past observations with known labels. In classification, the output variable is categorical or discrete, and the algorithm learns to map input features to

a specific class label. Common examples include logistic regression, decision trees, random forests, SVM, and neural networks [59].

A.1.2 Unsupervised Machine Learning

Unsupervised machine learning is defined as a ML model which learns from data without human supervision. The model is given unlabelled data from which it discovers patterns and insights without any explicit guidelines or instructions. Unsupervised learning algorithms are useful for identifying previously undetected patterns which can be useful in understanding and categorising data better [52].

A.1.2.1 Clustering

A common example of unsupervised learning is "clustering". Clustering algorithms such as k-means clustering can be used to group similar data points together without prior knowledge of class labels. The goal is to group similar data points together and discover underlying patterns or structures within the data [54]. Clustering can be used in image segmentation, market analysis, anomaly detection, document classification tasks and so on [55].

A.1.3 Semi-Supervised Machine Learning

Semi-supervised learning combines both supervised and unsupervised ML algorithms by using both labelled and unlabelled data to train AI models. These models are generally used for classification and regression tasks. Semi-supervised learning methods are especially relevant in situations where vast amounts of unlabelled data is easily acquirable in contrast to the availability of limited labelled data. In these cases, neither supervised nor unsupervised ML algorithms will provide adequate solutions [53]. In semi-supervised learning, techniques like self-training or co-training leverage a small amount of labeled data along with a larger pool of unlabeled data to improve model performance.

A.1.3.1 Self-training

Self-training employs a cyclical process in which a model is initially trained on labelled data, then used to confidently assign labels to additional unlabelled data points. These newly assigned labels are integrated into the training set for subsequent iterations, progressively enlarging the dataset and enhancing the model's accuracy. For instance, in natural language processing, a sentiment analysis model trained on a modest amount of labelled data can classify unlabelled text, using these confident predictions to expand the training set [60].

A.1.3.2 Co-training

In contrast, co-training leverages diverse perspectives or points of view of data to enhance learning. This method involves training several models on distinct feature subsets or data representations, each possessing its own labeled and unlabeled examples. Throughout the training phase, these models share insights by concurring

on the labels for unlabeled data points, thus strengthening their collective learning. Co-training is effectively used in text classification, where one model may evaluate the textual content of a document, while another concentrates on structural attributes like word frequency or syntax [61].

A.2 Chest X-Ray labels

A.2.1 Atelectasis

The collapse of a part of the lungs (a lobe) either due to blockage of the bronchus or bronchioles, or due to unexpected pressure on the lung, is termed "Atelectasis" (AT) [46].This condition occurs when the alveoli lose air. Symptoms of Atelectasis include difficulty in breathing, weak breathing, wheezing and coughing.

On X-rays, AT is typically seen as small volume linear shadows, often located peripherally or at the lung base. Lobar collapse presents with a more characteristic appearance corresponding to the affected lobe, while AT can exhibit a more varied position and appearance [45].

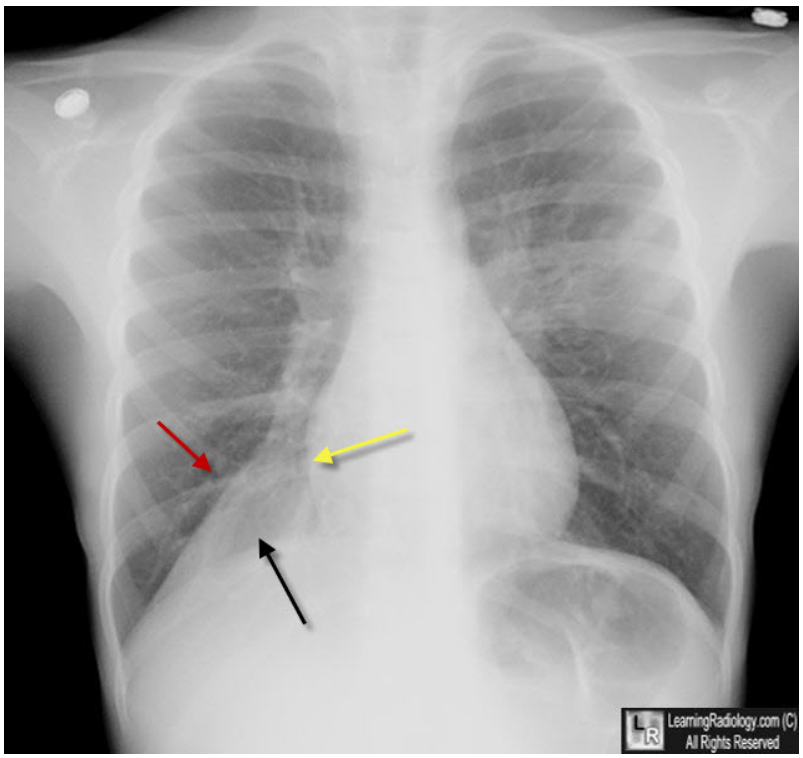


Figure A.1: Atelectasis, Right Lower Lobe. Source: [47]

Figure.A.1 illustrates the presence of AT in the right lower lobe, as indicated by the arrows. It appears as a darker shadow on the X-ray. Frontal chest X-rays are commonly utilized to detect Atelectasis (AT). Often, the underlying causes of AT,

such as tumors exerting pressure on the lobe or pleural effusion, can also be observed on the X-ray.

A.2.2 Pleural Effusion

Pleural effusion is the accumulation of excess fluid in the pleural cavity, the space between the two layers of the pleura that surround the lungs. This condition can be due to various underlying causes such as heart failure, infection, malignancy, or inflammatory diseases [100].

Pleural effusion commonly presents with several symptoms. These include shortness of breath, especially noticeable when lying down. Patients often experience chest pain, which tends to worsen with breathing or coughing, typically described as sharp. A persistent cough may accompany the condition, varying from dry to productive if there is an associated infection. Additionally, individuals may feel fatigued or generally weak

On X-rays, pleural effusion typically appears as a homogeneous, opaque area that obscures the costophrenic angle. It is often seen as a blunting of the angle between the diaphragm and the rib cage. Ultrasound can provide a more detailed visualization, showing the presence of fluid and its level within the pleural space. CT scans are useful for a more precise assessment and to identify the underlying cause.

A.2.3 Cardiomegaly

Cardiomegaly refers to an abnormal enlargement of the heart. It is often a secondary condition resulting from various underlying heart diseases such as hypertension, cardiomyopathy, heart valve disease, or congenital heart defects [101].

Individuals with cardiomegaly may experience a range of symptoms, including shortness of breath, particularly when exerting themselves or when lying flat. They may also feel fatigued and weak, and notice swelling in the legs, ankles, or abdomen. Additionally, palpitations or irregular heartbeats are common in those with this condition

Cardiomegaly is commonly identified on chest X-rays by an enlarged cardiac silhouette. The heart's width exceeds half of the thoracic diameter, and the heart's borders are indistinct. Echocardiography is the preferred method for detailed assessment, providing information on the heart's size, function, and structure. CT scans and MRI are also valuable for detailed imaging and evaluating heart chambers and surrounding structures.

A.3 Attributes and Correlations

For experiment 1, correlation between the attributes and the label: male is computed to assess the impact of the PI on the model.

Table A.3.1: Attribute Correlation

No.	Attribute	Correlation
1	Male	1.0
2	5 o' Clock Shadow	0.416
3	Big Nose	0.372
4	Wearing Necktie	0.329
5	Bags Under Eyes	0.301
6	Goatee	0.305
7	Sideburns	0.286
8	Mustache	0.242
9	Bushy Eyebrows	0.242
10	Chubby	0.231
11	Double Chin	0.207
12	Eyeglasses	0.200
13	Gray Hair	0.189
14	Bald	0.179
15	Wearing Hat	0.130
16	Receding Hairline	0.117
17	Black Hair	0.111
18	Big Lips	-0.168
19	Bangs	-0.158
20	Oval Face	-0.121
21	Wearing Necklace	-0.268
22	High Cheekbones	-0.247
23	Pointy Nose	-0.214
24	Rosy Cheeks	-0.212
25	Wearing Earrings	-0.373
26	Young	-0.291
27	Wavy Hair	-0.321
28	Blond Hair	-0.307
29	Attractive	-0.399
30	Arched Eyebrows	-0.407
31	Mouth Slightly Open	-0.099
32	Brown Hair	-0.109
33	Straight Hair	0.062
34	Pale Skin	-0.076
35	No Beard	-0.520
36	Smiling	-0.136
37	Heavy Makeup	-0.666

DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY