Blending Unsupervised and Supervised Mimic Learning for Discovering Interpretable Phenotypes in 3D Imaging Volumes

by

Mahboobeh Norouzi

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, Manitoba, Canada

Contents

List of Figures

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Dr. Ahmed Asharf, for his guidance, support, and encouragement throughout my thesis work. Dr. Ashraf's expertise, insightful comments, and constructive feedback were instrumental in shaping my research and improving my writing skills.

I would also like to thank Dr. Shehroz S. Khan for his valuable contributions to my thesis. His insightful suggestions have been invaluable in enhancing the quality of my research and strengthening my arguments.

Additionally, I am grateful to Dr. James Davie and Dr. Behzad Kordi for accepting to be my advisory committee members. Their input and feedback have helped me refine my ideas and enhance my work.

In the end, I want to acknowledge my greatest appreciation to my supportive family and friends throughout my academic journey. I am incredibly thankful to my parents for instilling a love for learning and always being there for me, regardless of the challenge. To my siblings, thank you for your constant love and support. I would also like to thank my friends for their constant support and for keeping me grounded during the ups and downs of my academic journey. Your encouragement and understanding helped me push through when times got tough.

DEDICATION

Dedicated to the brave and determined women of Iran, whose tireless quest for freedom inspires us all.

ABSTRACT

The ability to differentiate between different subtypes of diseases is essential for informed clinical decision-making. However, in cases where only partial knowledge of a disease exists and its subtypes are unknown, traditional supervised classification approaches are not applicable. Instead, unsupervised methods are necessary for subtype discovery. Previous research in imaging-based subtype discovery has been limited, often lacking interpretability in the discovered phenotypes.

This thesis proposes a novel data-driven method for discovering interpretable imaging phenotypes in 3D image volumes. While previous research has been limited, especially for 3D imaging modalities like MRI and CT, our approach focuses on CT scan images as a demonstration of 3D imaging. Our method combines unsupervised learning and supervised mimic learning for phenotype discovery and interpretability, respectively. We utilize unsupervised 3D autoencoders to discover phenotypes and employ supervised mimic learning to interpret our unsupervised pipeline. Notably, this is the first application of supervised mimic learning to understand an unsupervised model. Additionally, we introduce and formulate VolPAM, a technique enabling 3D interpretation of the discovered phenotypes. Our method is applicable to various medical imaging modalities and holds the potential to advance our understanding of respective diseases and conditions.

Chapter 1

Introduction

1.1 Motivation

The distinction between different subtypes of diseases plays a crucial role in clinical decision-making and patient management. However, in cases where our understanding of a disease is limited and its subtypes are unknown, traditional supervised classification approaches are not feasible. In such situations, unsupervised methods are required for subtype discovery. Previous research on discovering phenotypic subtypes based on medical imaging has been limited, often lacking interpretability in the discovered phenotypes [1, 2, 3, 4].

The supervised learning paradigm, particularly in deep learning, has been extensively employed in medical imaging for various classification tasks [5, 6, 7]. Deep learning techniques have shown remarkable performance in classifying different imaging modalities. However, the effectiveness of supervised learning heavily relies on the understanding and knowledge of imaging phenotypes, which can be challenging when subtypes of a particular condition are not well understood or constantly evolving. In such cases, the supervised paradigm may not effectively address the problem of disease subtype detection. Thus, there is a need for unsupervised data-driven approaches to discover intrinsic imaging phenotypes, allowing researchers to focus on specific elements of the identified phenotypes and advance their understanding of the disease.

1.2 Problem Definition

The data-driven discovery of imaging phenotypes, especially in 3D imaging modalities like computed tomography (CT) and magnetic resonance imaging (MRI), has received limited attention in situations where prior knowledge is lacking. Previous studies have focused on handcrafted features or specific imaging modalities [8, 9, 10, 11]. This thesis aims to address the problem of imaging phenotype discovery in a latent representation learned by unsupervised 3D autoencoders, departing from reliance on handcrafted features. Furthermore, the interpretation of the discovered phenotypes is challenging, and novel techniques are required to enhance interpretability and enable visualization in a 3D context.

1.3 Thesis Objectives

The objectives of this thesis are to develop a comprehensive data-driven methodology for discovering interpretable imaging phenotypes in 3D image volumes and enhance the interpretability of the discovered phenotypes. The proposed methodology will combine unsupervised and supervised learning paradigms. Specifically, the first objective is to leverage unsupervised techniques to learn representations that reveal intrinsic groupings or imaging phenotypes in 3D imaging volumes. The second objective is to use supervised learning to interpret the unsupervised model and gain a better understanding of the discovered phenotypes. The proposed methodology aims to provide valuable insights into medical conditions through the analysis of imaging phenotypes.

1.4 Methodology

The methodology begins by training a 3D convolutional autoencoder to learn a latent representation for 3D image volumes. The learned latent representation is then subjected to hierarchical clustering to identify unique phenotypes in the latent space. To interpret the identified phenotypes, supervised learning is employed to mimic the input-output characteristics of the unsupervised pipeline. Inspired by the concept of knowledge distillation, a 3D convolutional neural network (CNN) is trained to categorize the latent phenotypes using image volumes as input. Additionally, the

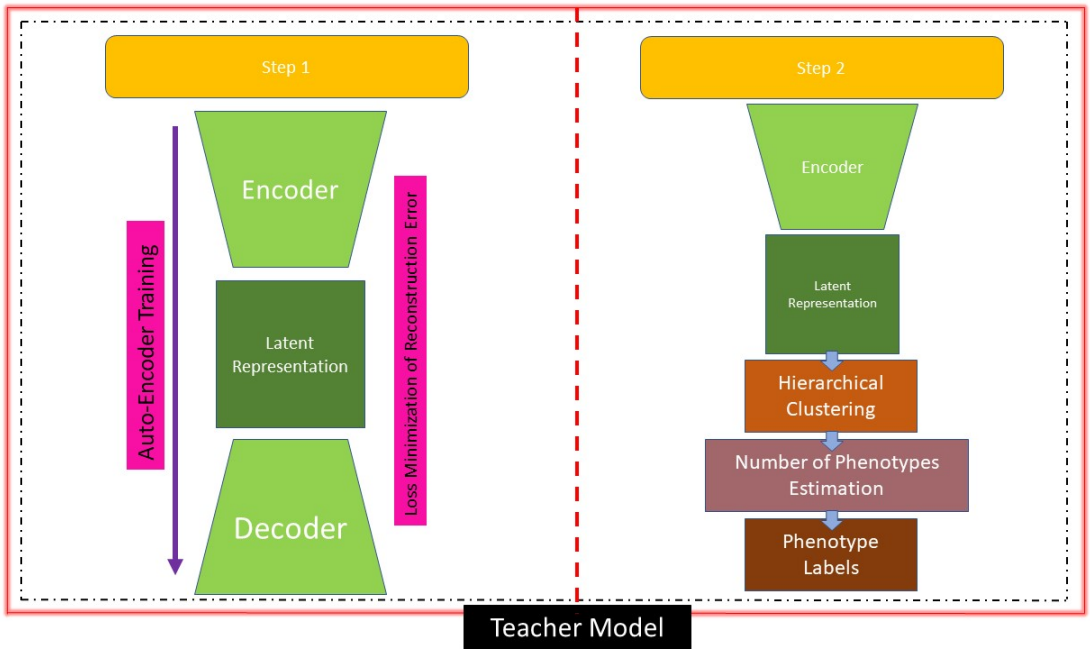
VolPAM technique is introduced to enhance the interpretability of the discovered phenotypes. VolPAM enables principled 3D visualization and interpretation of the phenotypes, identifying the voxels, 2D spatial regions, and slice depths contributing to a particular latent phenotype.

1.5 Contributions

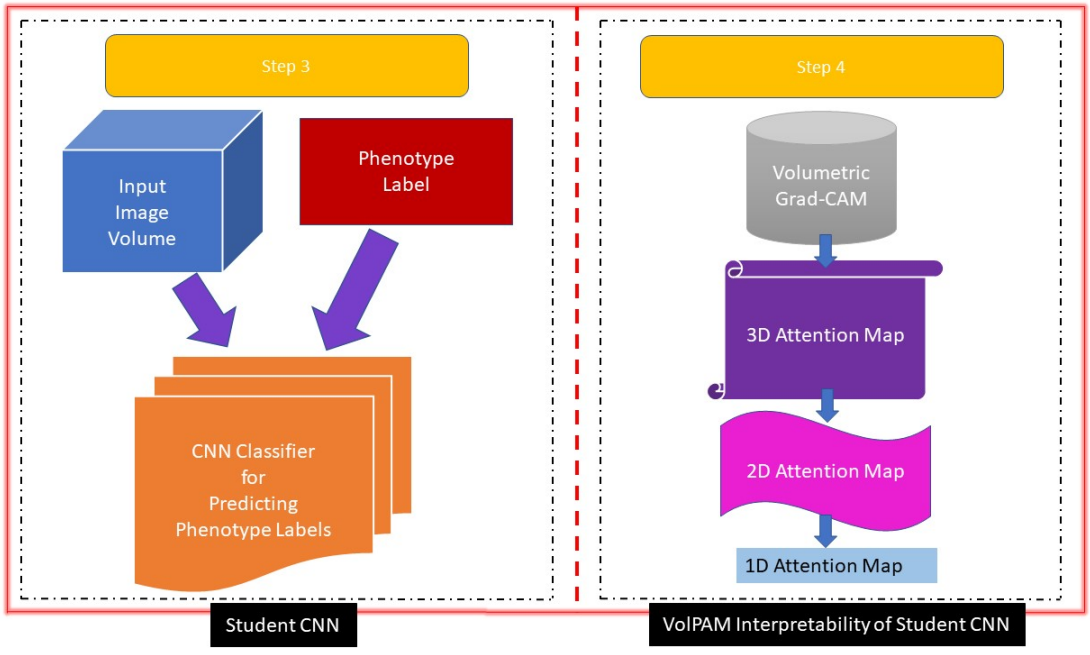
The contributions of this thesis include the development of a novel approach for discovering and interpreting imaging phenotypes. The proposed methodology combines unsupervised and supervised learning paradigms to discover intrinsic groupings in 3D imaging volumes and enhance the interpretability of the discovered phenotypes. The introduction of the VolPAM technique facilitates principled 3D visualization and interpretation of the phenotypes. These contributions have the potential to improve disease understanding and serve as valuable tools for clinicians and radiologists. The research conducted in this thesis has resulted in a manuscript entitled "VolPAM: Volumetric Phenotype-Activation-Maps for Data-driven Discovery of 3D Imaging Phenotypes and Interpretability" which has been submitted to the Neural Computing and Applications Journal and is currently under review [12].

1.6 Thesis Organization

The thesis is organized as follows. Chapter 2 provides a comprehensive review of relevant concepts and literature. Chapter 3 presents a detailed explanation of the methodology, including the training of the 3D convolutional autoencoder, hierarchical clustering, and the VolPAM technique. Chapter 4 presents the experimental results and evaluations of the proposed methodology on computed tomography (CT) scans, along with visualizations of the discovered phenotypes. Chapter 5 summarizes the findings, discusses their implications, and suggests potential future research directions. A graphical abstract of the central ideas of this thesis is shown in Figure 1.1:



(a)



(b)

Figure 1.1: Graphical abstract of the ideas presented in Chapters 3 and 4: (a) Chapter 3: Unsupervised Pipeline for Phenotype Discovery. The teacher model consists of the entire unsupervised pipeline (steps 1 and 2). (b) Chapter 4: Supervised Mimic Learning for Phenotype Interpretability.

Chapter 2

Literature Review

2.1 Artificial Intelligence and Machine Learning

The introduction of Artificial Intelligence (AI) in 1956 sparked the study of machines simulating intelligent human capabilities [13]. Machine Learning (ML), a subdiscipline of AI, focuses on building computer models that improve task performance through exposure to large datasets [14]. ML methods enable computers to learn rules and recognize patterns without explicit programming or deductive assumptions. By providing input data, ML models can be trained to produce desired results through optimization algorithms [5].

To ensure the effectiveness of ML, two prerequisites are crucial. First, relevant and detailed datasets must be available, offering sufficient information for the research question under focus[15]. Feature extraction and computation techniques enhance the datasets by deriving structured information representing specific attributes of data points [14]. The quality, precision, and richness of these features significantly impact the performance of learning algorithms. Improper or irrelevant data can lead to inadequate representation and hinder the development of useful models.

Considerable efforts have been devoted to feature extraction and selection techniques, addressing the importance of dataset representation [14]. Furthermore, deep learning approaches have demonstrated the capability to learn task-specific features, which will be further explored in later subsections.

In the following subsections, we will review important works in the field of ML, focusing on both supervised and unsupervised learning paradigms. These works have contributed to advancements in various applications and methodologies, providing

valuable insights for further research.

2.2 Supervised and Unsupervised Machine Learning

Supervised learning methods are more prevalent in the field of machine learning, where algorithms are trained on annotated datasets to estimate correct labels for new, unseen data [5]. The objective is to minimize the model's classification error by precisely computing output values based on input-output instances [16]. However, supervised learning heavily relies on manual labelling, which is expensive and challenging to scale, especially with the growing volume of visual data [17]. The limited availability of labelled data poses a significant constraint on the utility of machine learning algorithms. To address this limitation, there has been a growing interest in unsupervised learning, particularly in learning feature representations without manual annotation [18].

Unsupervised machine learning, in contrast to supervised approaches, focuses on discovering patterns in data without relying on annotation [5]. In recent years, deep convolutional neural networks have demonstrated their ability to learn high-level semantic information from images, revolutionizing computer vision. However, these networks typically require large amounts of labelled data, which is costly and challenging to obtain at scale [17]. Unsupervised semantic feature learning has emerged as a crucial approach to effectively leverage the unlabeled visual data which is available in abundance. By directly handling unlabeled input data, unsupervised algorithms alleviate the labour-intensive task of data labelling [19].

These advances in both supervised and unsupervised learning have significantly impacted various domains, including computer vision and pattern recognition. The exploration of novel algorithms and methodologies continues to enhance the capabilities of machine learning models in processing and understanding complex datasets.

2.3 Deep Learning

Deep learning (DL) has gained significant attention due to its ability to recognize complex patterns in large datasets [20]. DL models, known as deep artificial neural networks, have transformed the field of machine learning by automating the process of feature extraction [5]. Instead of relying on manually constructed features, deep learning models can learn representations directly from raw data, eliminating the

need for time-consuming feature engineering [5]. This capability to discover useful representations autonomously is a key characteristic of deep learning, setting it apart from traditional machine learning techniques [5].

The progress in deep learning has been driven by the availability of big data, user-friendly software frameworks, and increased computational power [5]. In the domain of computer vision, deep learning has shown superior performance in various image analysis benchmarks [5].

2.4 Artificial Neural Networks

Artificial neural networks (ANNs) were initially developed in the 1950s. They continue to remain a topic of wide research interest even today.[21]. ANNs are constructed with interconnected computational nodes, or neurons, which process input data through a series of layers [16]. The connections between neurons are associated with learnable weights, and the output of each neuron is calculated based on these weighted inputs [16]. ANNs can vary in depth, with deep neural networks consisting of multiple layers [21]. Training ANNs involve optimizing the network parameters to minimize a defined loss function, typically using gradient-based methods [22, 23]. The goal is to find parameter values that reduce the discrepancy between the ANN's output and the desired target [22]. Different variants of gradient descent algorithms exist to improve convergence properties [23]. While fully connected networks (FCNs) were initially prevalent, specialized network architectures have been proposed for specific data formats. Convolutional neural networks (CNNs) have been particularly successful in capturing spatial patterns in images [24].

Figure 2.1 illustrates an example of an FCN architecture, which consists of an input layer, multiple hidden layers, and an output layer. However, the suitability of FCNs depends on the type of input data, and other network architectures have been developed to address specific data formats such as images, video, time series, and signals [20].

2.5 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are widely used for imaging and computer vision tasks. They gained popularity after achieving state-of-the-art results in the 2012 ILSVRC object recognition competition [25, 26]. CNNs have since been applied in

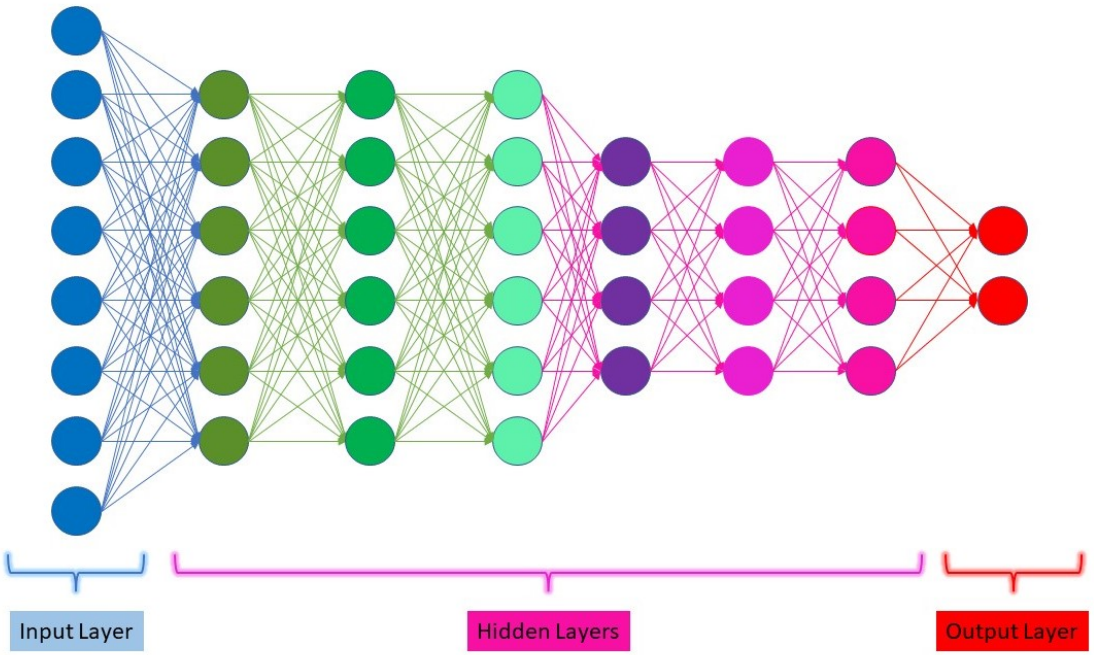


Figure 2.1: A fully connected neural network architecture consisting of an input layer, six hidden layers, and one output layer. The input layer consists of 8 nodes, followed by three hidden layers with six nodes each and three hidden layers with four nodes each. The final output layer contains two nodes representing the network's predicted outputs.

various applications, including facial recognition, self-driving cars, unmanned grocery stores, and intelligent medical treatments [27].

The fundamental idea of CNNs is to compute convolutions over local neighbourhoods, allowing them to capture different aspects of the input. This local connectivity reduces the number of parameters compared to fully connected networks [16]. CNNs can handle multi-channel inputs, such as RGB or hyperspectral images, as well as 3D imaging volumes [16]. A typical CNN architecture consists of alternating convolutional layers and pooling layers for dimensionality reduction, followed by fully connected layers that produce learned feature representations. Figure 2.2 illustrates the architecture of a typical CNN.

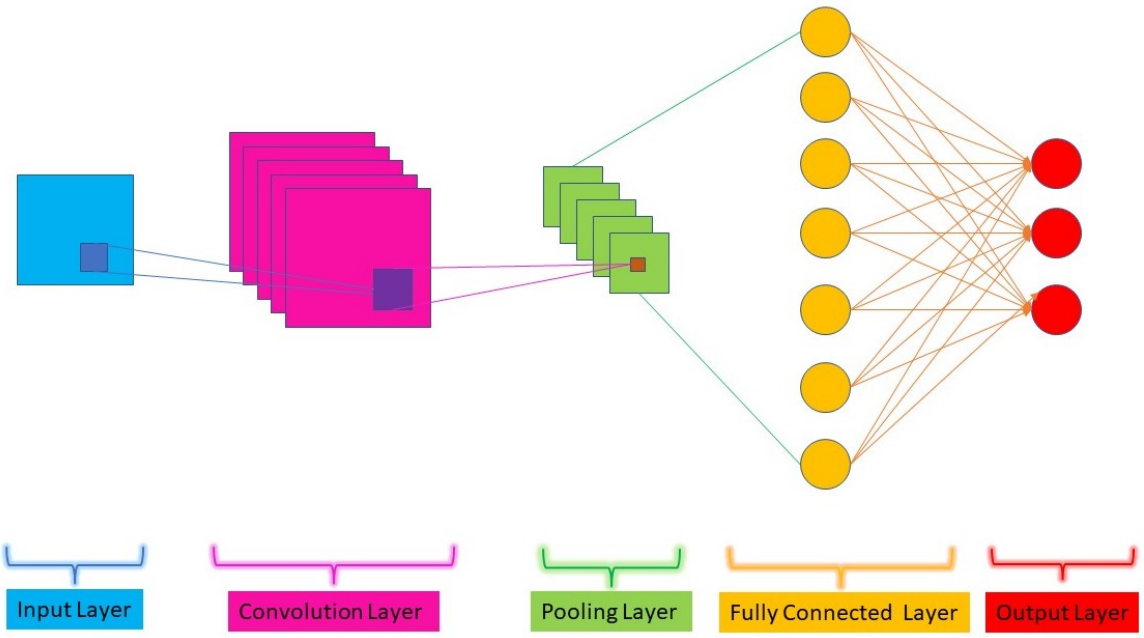


Figure 2.2: A typical Convolutional Neural Network (CNN) architecture with an n x n input layer, a c x c convolutional layer with a depth of 5, a p x p pooling layer with a depth of 5, a fully connected layer with six nodes, and an output layer with three nodes. The convolutional layer applies filters to extract features from the input data, followed by the pooling layer to reduce dimensionality. The fully connected layer connects all nodes to the next layer, leading to the final output layer with three nodes representing the predicted classes or values.

2.6 Clustering

Clustering is a sub-category of unsupervised machine learning that aims to classify unlabeled data into groups based on uniform patterns. It is useful for organizing information and understanding relationships within data [28]. Traditional clustering approaches involve extracting feature vectors from raw data and applying clustering techniques to these vectors [29]. Deep learning techniques have been integrated with clustering, enabling direct clustering of original images with improved performance [29].

Clustering algorithms can be categorized into partitional and hierarchical methods. Partitional clustering creates non-overlapping clusters, such as the widely used K-means algorithm [30]. Hierarchical clustering generates nested clusters in a tree-like structure known as a dendrogram [31]. Agglomerative clustering merges similar clusters in a bottom-up approach, while divisive clustering divides clusters in a top-down manner [32]. Dendrograms provide visualizations of the clustering process [33].

Clustering has been applied in medical imaging to identify disease phenotypes and understand the relationship between clinical data and imaging [14].

2.7 Latent Representation Learning

Auto-encoders are neural networks used for unsupervised learning and learning useful representations of input data [34, 35, 36, 37]. They can compress data, improve computation speed, and enhance performance [38]. Auto-encoders consist of an encoder-decoder architecture where the encoder transforms the input into a latent representation, and the decoder reconstructs the input from the latent representation [18]. The encoder is also used as a feature extractor [18]. Auto-encoders can be designed with different architectures depending on the data format, such as fully connected layers for simple feature vectors or convolutional layers for images [39]. Convolutional Auto-Encoders (CAEs) and Deep Auto-Encoders (DAEs) are examples of auto-encoder variants [39, 40, 41].

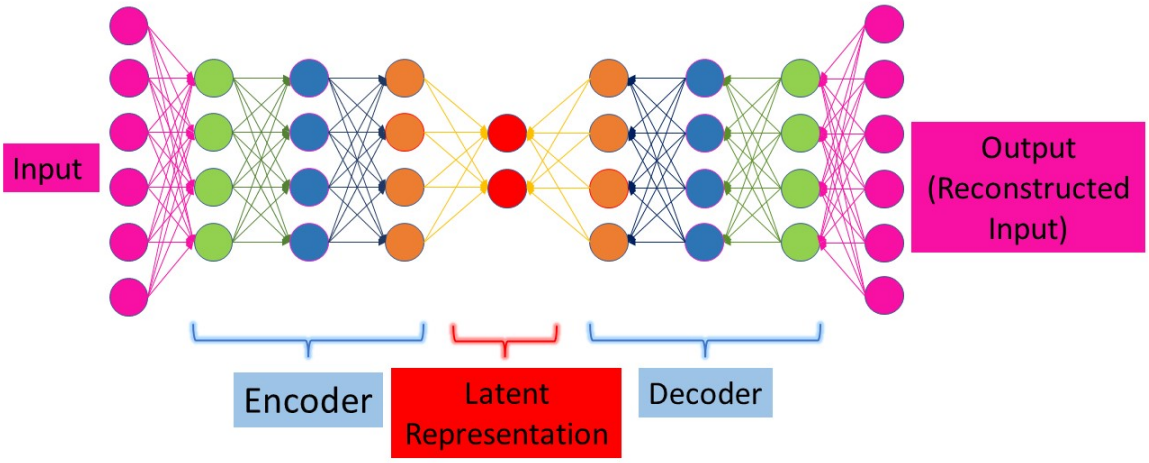


Figure 2.3: A typical autoencoder architecture with an input layer, three encoding layers, one latent representation layer, three decoding layers, and one output layer. The input will be reconstructed through encoder and decoder layers. Throughout this process, the hidden (latent) representation learns useful information about the data in a lower dimension.

2.8 Neural Network Interpretability

Interpreting neural networks is essential for understanding their decisions. There are two high-level categories of interpretability methods. The first category is intrinsic interpretability which involves building self-explanatory models, such as decision trees and rule-based systems [42]. The second category is Post-hoc interpretability techniques that provide explanations after the model has been trained. Examples include Local Interpretable Model Agnostic Explanations (LIME) and Shapley Additive exPlanations (SHAP) [43, 44]. Attention-based methods focus on salient regions of an image, either through learnable attention during training or post-hoc attention identification [45]. Grad-CAM is a popular attention-based method that provides visual explanations for CNN-based networks [45]. Knowledge distillation through mimic learning is another post-hoc interpretability approach, where a smaller, interpretable model mimics a larger, complex model [46].

2.9 CT Scans and AI in Medical Imaging

CT scans are widely used in medical imaging, particularly for pulmonary disorders. They are essentially acquired by performing multiple X-rays from different viewing angles, and then they can be algorithmically reconstructed to build 3d imaging volumes [47, 48, 49, 50]. AI, especially deep learning, has shown promise in diagnostic imaging tasks [51, 52, 53, 54, 55, 56, 57]. Large labelled datasets and advancements in GPUs have facilitated the success of AI in various medical imaging modalities, including CT [58, 20, 59, 26]. The integration of AI methods in radiology is expected to increase to address the challenges posed by the large volume of medical image data and the increasing workload demands on radiologists [60, 61, 53]. However, the success of AI in imaging has been primarily in the supervised paradigm with access to annotated datasets. Unsupervised AI methods are needed to handle situations where disease subtypes are not fully understood and annotated datasets are lacking.

2.10 Unsupervised Phenotype Discovery Literature Review

In this literature review, we will review some of the representative studies in the domain of unsupervised discovery of phenotypes.

One pivotal study utilized unsupervised hierarchical clustering to identify distinct imaging phenotypes in estrogen receptor-positive breast cancer tumours [62]. Breast cancer, a complex disease with varied clinical outcomes, necessitates reliable prognostic markers. This paper presents a method using dynamic contrast material-enhanced (DCE) magnetic resonance imaging (MRI) data to identify intrinsic imaging phenotypes in estrogen receptor-positive breast cancer tumours. The study aims to determine associations between these phenotypes and prognostic gene expression profiles, shedding light on their potential as imaging biomarkers to guide treatment decisions. The authors retrospectively analyzed DCE MRIs from 56 women diagnosed with estrogen receptor-positive breast cancer between 2005 and 2010. The primary tumours were assayed using a validated gene expression assay to assess the likelihood of recurrence. A multiparametric imaging phenotype vector was extracted for each tumour, encompassing quantitative morphologic, kinetic, and spatial heterogeneity features. Unsupervised hierarchical clustering was employed to identify distinct imaging phe-

notypes. The study revealed the presence of four dominant imaging phenotypes. Notably, two phenotypes comprised only low- and medium-risk tumours, signifying a potential association between imaging phenotype and recurrence likelihood. The analysis demonstrated that certain imaging features were correlated with the risk of recurrence. For instance, tumours with higher proportions of pixels representing the most enhancing areas and higher mean peak enhancements were linked to a higher risk of recurrence. Conversely, tumours with higher proportions of intermediate enhancing pixels and greater variances in wash-in slope exhibited lower risks of recurrence. By using multiparametric MRI data, physicians can determine the imaging phenotype of a breast cancer tumour, allowing for tailored approaches to patient care. Additionally, the correlations between imaging phenotypes and gene expression profiles provide a deeper understanding of breast cancer biology and its heterogeneity, potentially guiding further research and targeted therapies. Beyond recurrence likelihood, these imaging features could be explored for their associations with other clinical outcomes, such as treatment response and overall survival. Although the integration of unsupervised hierarchical clustering with DCE MRI data has proven to be a valuable tool for identifying intrinsic imaging phenotypes, there are a number of limitations in this study. The most important shortcoming is that the features used in this study are a hand-crafted set of features from DCE MRI on which the hierarchical clustering is being run. One particular problem with manually defined features is the inadvertent inclusion of human bias during the formulation of these features. The above reviewed study does not capitalize on the capability of deep learning to learn the features in a data-driven fashion using end-to-end training.

The second study that we are going to review is based on a Bayesian approach along with Monte Carlo dropout for deep neural networks to define uncertainty measures for each phenotype prediction [9]. Utilizing high-content screening (HCS) data, they successfully discover phenotypes based on these uncertainty measures. In real-world HCS experiments, it is often challenging to anticipate all phenotypes in advance, making the discovery of unknown phenotypes a critical requirement. Traditional deep learning approaches are limited in addressing this aspect of HCS. To tackle this challenge and fulfill the need for novelty detection, the authors employ the Monte Carlo dropout approach, defining different uncertainty measures for each phenotype prediction. Through the use of HCS data, these uncertainty measures prove effective in identifying hitherto unclear phenotypes. Additionally, the MC dropout method significantly improves classification accuracy, offering potential benefits for existing

network architectures. Subtle changes in variations in cell lines can lead to the emergence of phenotypes that were not part of the training data. The authors employ a convolutional neural network (CNN) architecture and a cell painting assay to analyze compartment-specific labelled proteins. Four uncertainty estimators derived from MC dropout predictions are tested, with three successfully detecting unseen classes. By utilizing MC dropout during testing time, the approach achieves accurate phenotype classification and effective novelty detection based on uncertainty measures. Experimental tests on yeast datasets and other datasets consistently demonstrate the discrimination between known and unknown classes during testing time. Moreover, the paper discusses the challenges associated with using deep learning to analyze HCS data, which involves analyzing a vast number of images to identify phenotypic endpoints accurately. To address this, the authors propose a method using Monte Carlo dropout to quantify uncertainty measures in CNN-based predictions, enabling the detection of novel phenotypes not covered by known training classes. This approach was demonstrated using a dataset of budding yeast cells with 19 subcellular localization classes. The shallow CNN model with dropout and data augmentation was evaluated through experiments where one or two localization classes are removed from the training and validation sets. The results demonstrate the method's accuracy in detecting novel phenotypes and providing a measure of uncertainty in predicted probabilities.

Another study that discovers phenotypes in an unsupervised manner on HCS data introduces a solution to the limitations of supervised machine learning in analyzing HCS data [8]. Their approach presents the CellCognition Explorer framework. CellCognition Explorer's methods enable unsupervised classification tasks, allowing the software to yield phenotype scores based on the deviation of cell morphologies from negative control images. The software's versatility is highlighted through its successful application to various large-scale screening datasets concerning nuclear and mitotic cell morphologies. By discovering rare phenotypes without prior user training, CellCognition Explorer proves to have broad implications for improved assay development in HCS. The authors present a method for screening cellular phenotypes using deep learning and novelty detection techniques. To evaluate their approach, they conducted experiments on HeLa cells expressing a fluorescent histone marker and RNAi depletion of 1214 genes previously identified as important for mitosis from a genome-wide screen. The researchers manually annotated 10,000 cell objects into nine different morphology classes and established a reference annotation using con-

ventional numerical features. Subsequently, they tested the performance of different novelty detection methods and compared phenotype scoring with conventional supervised learning. Notably, their method successfully identified the top 100 ranked siRNA hits previously identified by conventional supervised learning.

Although the above methods have been a significant advance, they are primarily focused on HCS screening data. It has been shown that subtypes and molecular subtypes of different conditions have imaging footprints in 3D imaging modalities such as CT and MRI. As there is limited work in this domain, there is a need to develop methods for discovering phenotypes from 3D imaging volumes, which will be the focus of this thesis.

2.11 Summary

A comprehensive overview of relevant concepts in machine learning and artificial intelligence was provided in this chapter. The distinction between supervised and unsupervised learning was discussed, emphasizing the importance of labelled and unlabeled data in training machine learning models. Deep learning, specifically convolutional neural networks (CNNs), was introduced as a powerful approach for learning hierarchical representations from raw data, particularly in the context of image analysis. Clustering methods for discovering patterns in unlabeled data were explored, along with the concept of latent representation learning using autoencoders. The significance of neural network interpretability and various techniques for understanding model decisions were also addressed. Furthermore, the role of AI, specifically deep learning, in medical imaging, with a focus on CT scans, was examined. Also, a review of three key studies for the unsupervised discovery of phenotypes has been presented. This chapter served as the foundation for subsequent chapters, which delve into unsupervised representation learning and phenotype discovery.

The following chapter will focus on the first part of the work, which is the unsupervised discovery of the phenotypes.

Chapter 3

Unsupervised Phenotype Discovery

In this chapter, the first part of the work which is the unsupervised discovery of imaging phenotypes will be introduced and discussed and the results will be shown.

3.1 Datasets

Even though the techniques outlined in this study can be used with any 3D imaging modality, we have evaluated our algorithms on the following two datasets:

1. **The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) [63]:** LIDC/IDRI is a thorough reference database on lung nodules. Volumes of helical thoracic CT images from 1010 people are stored in this database. With the appropriate local Institutional Review Board (IRB) approvals, the photos were retroactively obtained from the picture archiving and communications systems (PACS) of seven cooperating academic institutions.

2. **COVID19-CT-Dataset [64]:** This publicly-accessible database consists of 1009 chest CT images of patients with a confirmed COVID-19 diagnosis. CT scans were gathered from two significant general university hospitals in Mashhad, Iran, between March 2020 and January 2021. All information is saved in DICOM format and comprises 512×512 pixel 16-bit grayscale images. All CT scans were examined visually by two board-certified radiologists to be checked for COVID-19 infections. A more experienced radiologist made the final call in the case that the two radiologists could not agree. A variety of COVID-specific lung infection patterns were visible on CT scans.

3.2 Representation Learning

One objective of this study is to discover phenotypes without the need for supervision. Therefore, the choice of representation is crucial for this purpose. Using raw image volumes can be problematic since, in that case, every voxel is treated as a feature, and the resulting phenotypes may be vulnerable to noise and variability due to translation. In recent years, deep autoencoders have become a popular method for unsupervised representation learning [35, 36, 37].

The specific architecture of the encoder and decoder in an autoencoder depends on the input's format, such as CNNs for images or a recurrent neural network for sequences [65]. A neural network specifically created for image processing tasks like compression and denoising is known as a convolutional autoencoder. By using convolutional layers rather than fully connected layers, it improves the fundamental autoencoder structure. Convolutional layers are used as a learnable downsampling step to encode the input image into a compact representation, and then transposed convolutional layers are used as a learnable upsampling step to decode the compact representation back into the original input image. The main goal is to minimize the difference between the input and output images while also minimizing the compressed representation's size. Contrary to fully connected layers, convolutional layers enable more effective and efficient extraction of picture features [39].

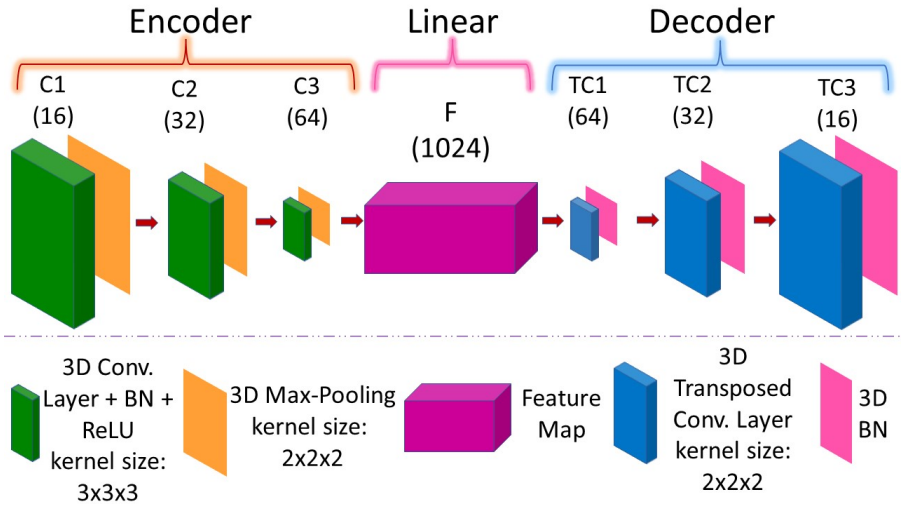
The experimental data consists of 3D image volumes, so we employed a 3D convolutional autoencoder (CAE) with a 3D CNN for both the encoder and decoder. The 3D CAE's architecture is illustrated in Figure 3.1. The architecture of our CAE comprises three convolutional layers with 16, 32, and 64 kernels. The diversity of the learned features is dependent on the number of kernels in each convolutional layer. The CAE can capture a variety of properties at various levels of abstraction by utilizing several kernels. Learning hierarchical representations of the input images can be facilitated by starting with fewer kernels and gradually expanding them [29]. The output of the last layer in the encoder is then flattened and fed through a linear layer with 1024 features. As a bottleneck layer, the linear layer with 1024 features lowers the latent representation's dimensionality. This information compression aids in capturing the most crucial details while omitting unnecessary features [66]. The latent representation of CAE with the size of a size of (batch, 1024) is subsequently reshaped into a volume and run through convolutional layers with 64, 32, and 16 kernels, respectively, to reconstruct the initial image volume. The reconstruction layers'

selection of 64, 32, and 16 kernels mirrors the CAE's encoding stage. The decoder reconstructs the initial image volume in reverse by gradually reducing the number of kernels. Our CAE model has been trained with a batch size of 10, a learning rate of 0.001, and an RMSProp optimizer for 500 epochs. The CAE optimization process is affected by the batch size of 10 employed for training. Although a bigger batch size necessitates more memory, it can result in more stable gradients and quicker convergence. The batch size has been chosen based on the computational resources available and the trade-off between convergence speed and memory restrictions [66]. Hyperparameters that affect training include the RMSProp optimizer and the learning rate of 0.001. The gradient descent step size is determined by the learning rate, which also influences the likelihood of being stuck in local optima. The training efficiency and stability can be increased using the RMSProp optimizer, which adjusts the learning rate based on the size of recent gradients [67].

3.3 Phenotype Discovery

To identify inherent groupings and image phenotypes within a given dataset, clustering techniques are utilized on the latent representations obtained through the Convolutional Autoencoder (CAE). The goal of clustering is to divide a finite set of objects into groups, where similar objects are assigned to the same cluster and dissimilar objects to different clusters [68, 69]. Clustering algorithms are broadly categorized as partitional or hierarchical [70]. Partitional techniques such as K-means aim to minimize a loss function (e.g. within-group variance) while assigning data points to non-overlapping clusters. Probabilistic extensions of K-means, such as fuzzy c-means (FCM), allow for assigning data points to clusters with a probability distribution [71]. Hierarchical clustering, on the other hand, is more effective for discovering inherent groupings as it allows the data points to be arranged into nested clusters, forming a hierarchy of cluster membership that can be explored at multiple levels [62]. Hierarchical clustering can be performed in either a top-down (divisive) or bottom-up (agglomerative) approach [32].

To perform agglomerative hierarchical clustering, we utilized the latent representations of the data points obtained through the CAE. Initially, each data point was considered to belong to an individual cluster. At each iteration, the two closest clusters were merged, and the cluster centroids were recomputed using the Euclidean distance between the cluster centroids represented in the latent autoencoder space.



(a)

Layer (type)	Output Shape	Param #
Conv3d-1	[-1, 16, 128, 128, 64]	432
BatchNorm3d-2	[-1, 16, 128, 128, 64]	32
ReLU-3	[-1, 16, 128, 128, 64]	0
BasicConv-4	[-1, 16, 128, 128, 64]	0
MaxPool3d-5	[-1, 16, 64, 64, 32]	0
Conv3d-6	[-1, 32, 64, 64, 32]	13,824
BatchNorm3d-7	[-1, 32, 64, 64, 32]	64
ReLU-8	[-1, 32, 64, 64, 32]	0
BasicConv-9	[-1, 32, 64, 64, 32]	0
MaxPool3d-10	[-1, 32, 32, 32, 16]	0
Conv3d-11	[-1, 64, 32, 32, 16]	55,296
BatchNorm3d-12	[-1, 64, 32, 32, 16]	128
ReLU-13	[-1, 64, 32, 32, 16]	0
BasicConv-14	[-1, 64, 32, 32, 16]	0
MaxPool3d-15	[-1, 64, 16, 16, 8]	0
Flatten-16	[-1, 131072]	0
Linear-17	[-1, 1024]	134,218,752
Linear-18	[-1, 131072]	134,348,800
UnFlatten-19	[-1, 64, 16, 16, 8]	0
ConvTranspose3d-20	[-1, 32, 32, 32, 16]	16,416
BatchNorm3d-21	[-1, 32, 32, 32, 16]	64
ConvTranspose3d-22	[-1, 16, 64, 64, 32]	4,112
BatchNorm3d-23	[-1, 16, 64, 64, 32]	32
ConvTranspose3d-24	[-1, 1, 128, 128, 64]	129

Total params: 268,658,081
Trainable params: 268,658,081
Non-trainable params: 0

(b)

Figure 3.1: (a) The architecture of CAE. (b) The different layers of CAE, their respective output shapes, and the number of trainable parameters are presented in the model summary.

The optimal number of phenotypes was identified using the elbow method [72], where the average intra-cluster distance was computed and plotted against the number of clusters. The elbow method identifies the point on the curve corresponding to the highest percentage reduction in D_{avg} as the optimal number of clusters [62]. The average intra-cluster distance (ICD) was calculated using the following formula:

$$D_{\text{avg}}(k) = \frac{1}{k} \sum_{l=1}^{k} \sqrt{\frac{\sum_{u=1}^{N_l} \sum_{v=1}^{N_l} |X_{ul} - X_{vl}|^2}{N_l \left(N_l - 1\right)}} \tag{3.1}$$

where k is the number of clusters, and the letter l denotes the cluster index. The quantity N_l represents the number of data points in the l-th cluster, and X_{ul}, X_{vl} are the feature vectors for the u-th and v-th data points in cluster l [62]. The optimal number of distinct clusters, k^*, is determined using the following equation:

$$k^* = \arg\max_{k} \left(\frac{D_{\text{avg}}(k-1) - D_{\text{avg}}(k)}{D_{\text{avg}}(k-1)} \right) \tag{3.2}$$

3.4 Results

The proposed approach utilizes representation learning through an autoencoder followed by hierarchical clustering in the learned latent space. The dendrogram for the hierarchical clustering procedure for each dataset is shown in Figure 3.2. The optimal number of phenotypes (k^*) for the LIDC/IDRI and COVID19-CT datasets were identified using 3.2 and found them to be two and three, respectively. The dotted horizontal line in each dendrogram shows the level corresponding to two and three phenotypes. For better visualization, the dendrograms have been chopped at an earlier stage instead of going down to the bottommost level, where each cluster has one data point. As such, the numbers written in parentheses under the branch corresponding to each cluster represent the number of examples in the cluster at that level of the clustering hierarchy. For the LIDC dataset, out of 1010 data points, 606 were assigned to Phenotype 0, while 404 were assigned to Phenotype 1. In the Covid19-CT dataset, out of 1009 datapoints, 492, 296, and 221 were assigned to phenotypes 0, 1, and 2, respectively.

For robustness analysis, ten trials of autoencoder training were conducted, and clustering was performed over the learned latent representation each time. Figures 3.3 and 3.4 show the variation of intra-cluster distances as a function of the number

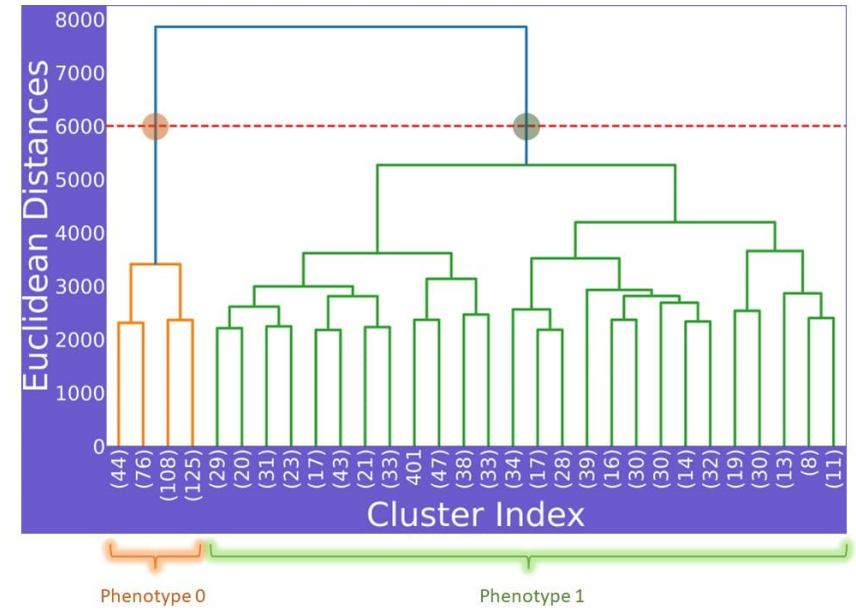
of clusters over ten trials for each dataset. Equation 3.2 was used to estimate the optimal number of phenotypes. For the LIDC dataset (Figure 3.3), across all ten trials, the optimal number of phenotypes, k^* was found to be two i.e. $k^* = 2$ resulted in the greatest percentage reduction in the average intra-cluster distance (D_{avg}). For the Covid19-CT dataset (Figure 3.4), for 8 out of 10 trials, the optimal number of phenotypes was found to be three.

So far in the thesis, the unsupervised representation learning followed by hierarchical clustering along with the analysis of intra-cluster distances has been shown, which can allow the discovery of an optimal number of clusters in the latent space. In the next chapter, we focus on how supervised learning can be leveraged to interpret the discovered imaging phenotypes.

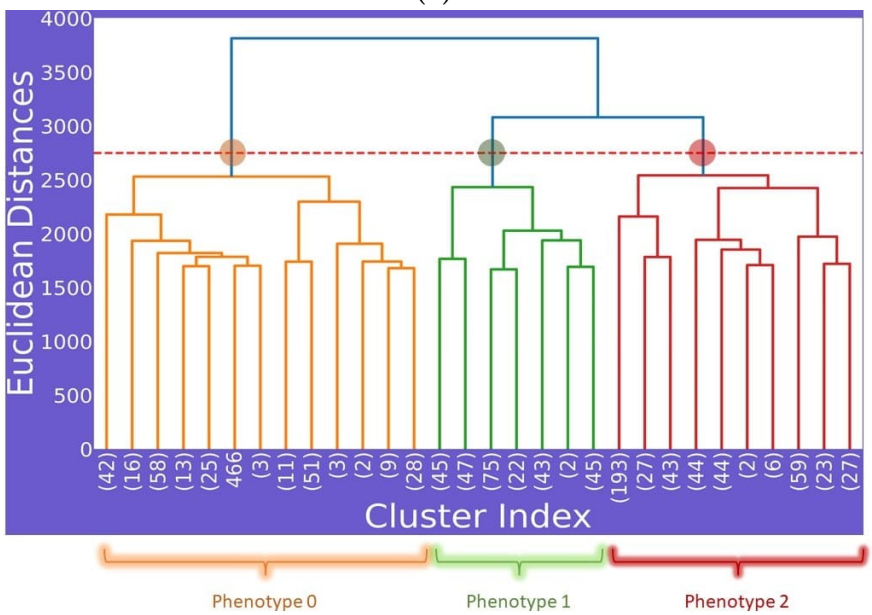
3.5 Summary

The process of discovering phenotypes without supervision was explored in this chapter. The datasets used, including LIDC, IDRI, and the COVID19-CT Dataset, were introduced. The concept of unsupervised representation learning was discussed, specifically focusing on 3D convolutional autoencoders (CAE) as a method for learning representations. The architecture of the CAE, emphasizing convolutional layers for feature extraction, was described. Phenotype discovery was then addressed, utilizing clustering algorithms on the latent representations obtained from the CAE. The agglomerative hierarchical clustering approach was employed to reveal inherent groupings and phenotypes. The optimal number of phenotypes was determined using the elbow method. Results demonstrated successful phenotype discovery in both datasets, with robustness analysis performed. The chapter concluded by highlighting the importance of supervised learning in interpreting the discovered imaging phenotypes, to be further explored in the next chapter.

Our experiments have been performed using NVIDIA RTX A4000 GPU. Using this server as our resource, this part of our work took one day of processing and training for each dataset. The most part of this time has been used to train the autoencoder.



(a)



(b)

Figure 3.2: The hierarchical clustering analysis conducted on the LIDC and Covid19-CT datasets revealed two and three distinct phenotypes, respectively, as illustrated in the corresponding plots.

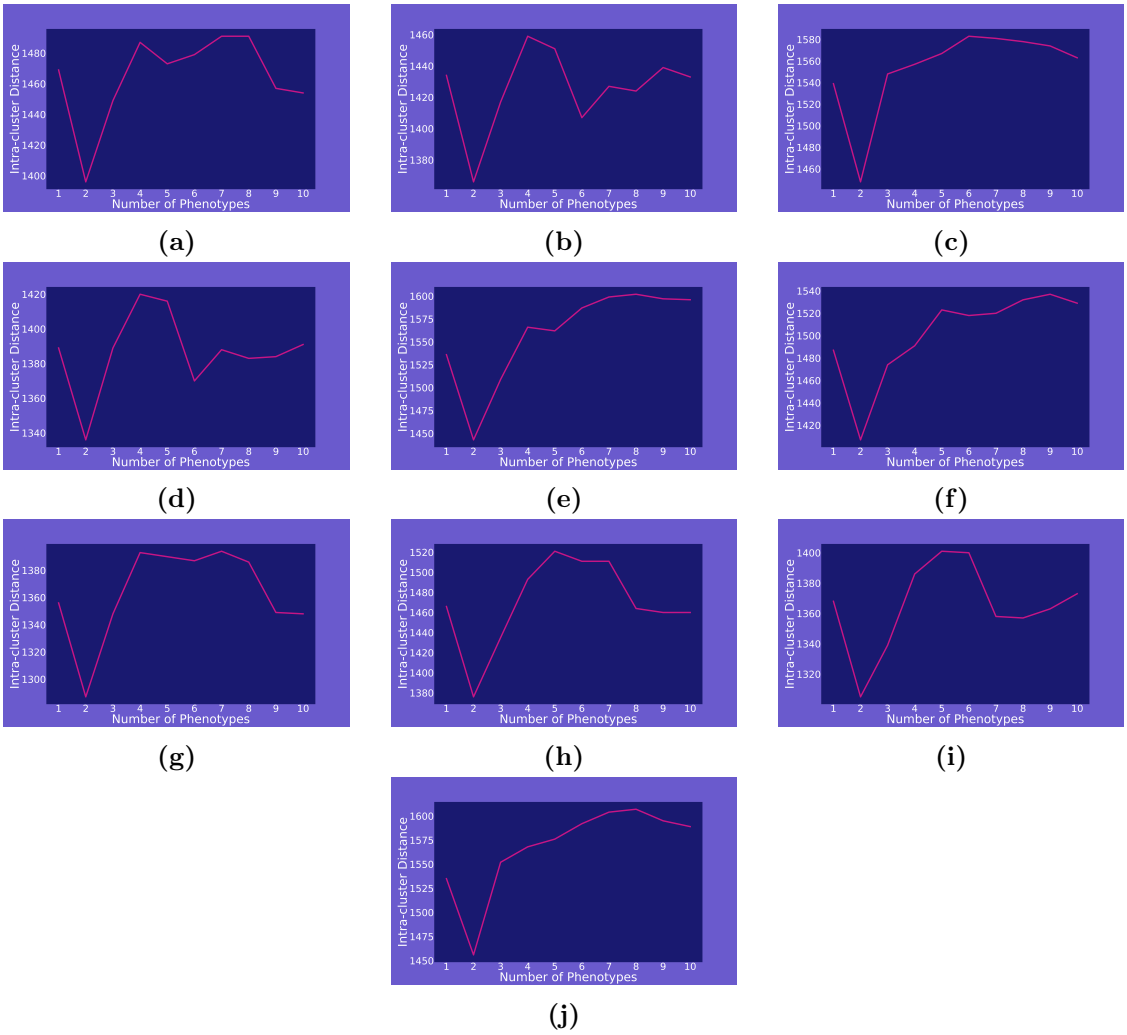


Figure 3.3: 10 plots depicting the intra-cluster distances for the LIDC dataset, calculated in the latent representation as a function of the number of phenotypes.

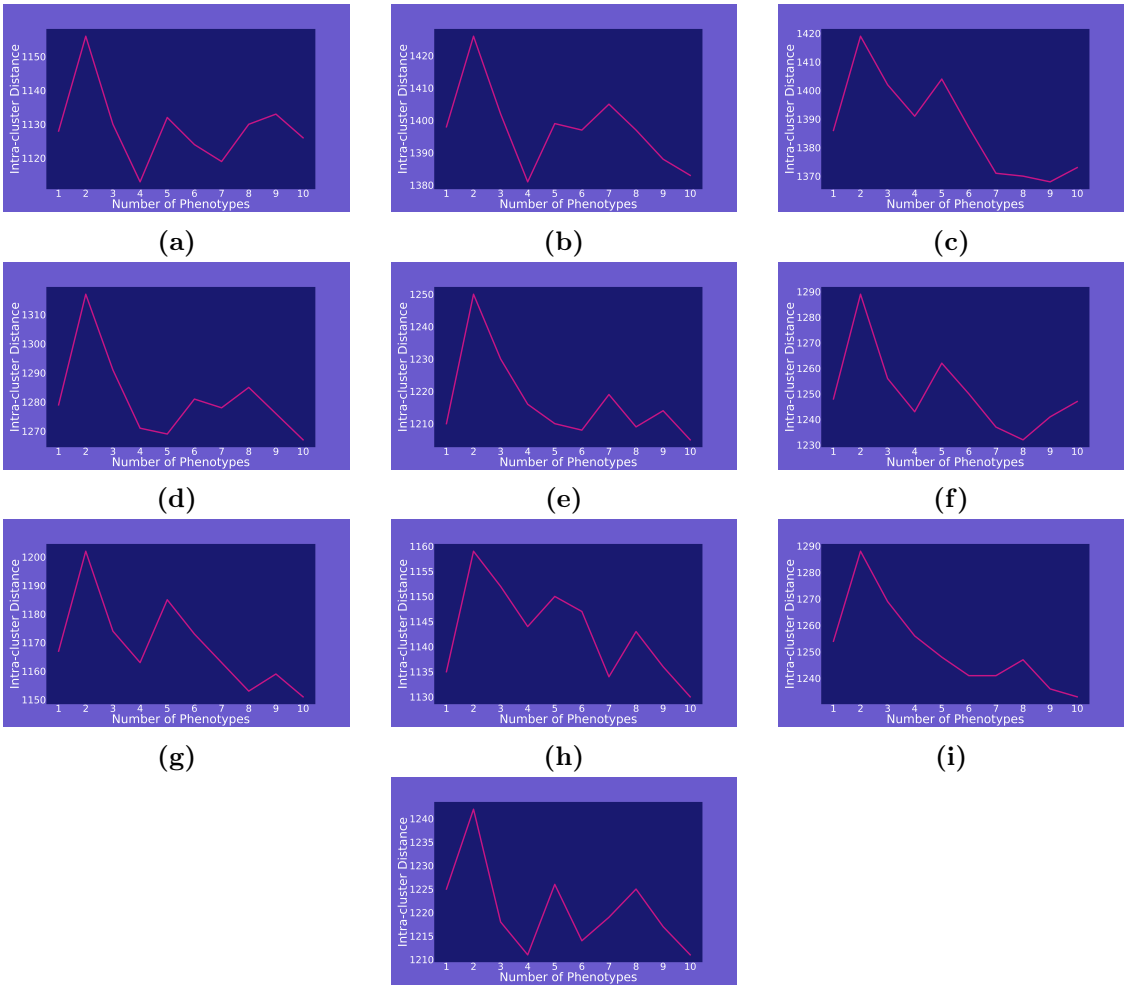


Figure 3.4: 10 plots depicting the intra-cluster distances for the Covid19-CT data, computed in the latent representation, plotted as a function of the number of phenotypes.

Chapter 4

Phenotype Interpretability through Supervised Mimic Learning

4.1 Phenotype Interpretability

In the previous chapter, a method for discovering phenotypes through unsupervised representation learning followed by hierarchical clustering was described. In this chapter, a general data-driven approach for interpreting these phenotypes will be proposed. Our goal is to interpret the unsupervised approach presented in the last chapter and decipher what parts of the image volume in terms of 3D, 2D, and 1D (along the depth) contribute toward each phenotype. Toward this goal, the key idea is to make use of supervised learning to train a model for classifying the phenotype label that the unsupervised approach of the previous chapter would have assigned to an input image volume. The phenotype labels from the unsupervised approach will act as ground-truth labels for this learning step. Through this step, essentially training a neural network in a supervised manner that learns to "mimic" the entire unsupervised pipeline, including representation learning, hierarchical clustering, and assigning clustering labels, is being proposed.

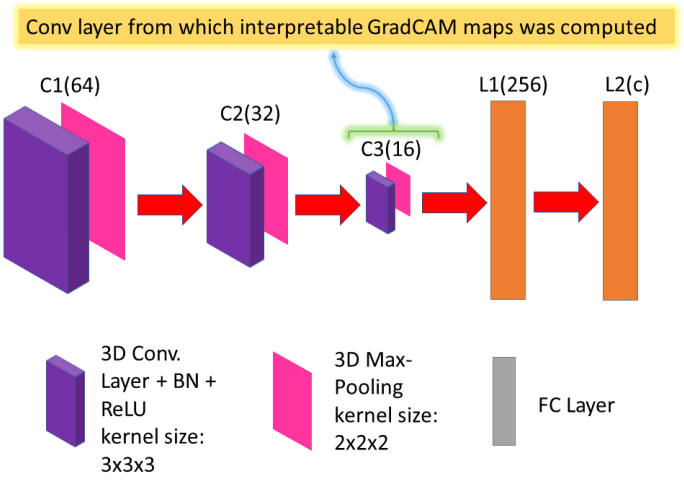
The motivation is that once a supervised model exists that can mimic the unsupervised pipeline, one can leverage the interpretability methods that have been recently developed to interpret neural networks trained in the supervised paradigm, particularly convolutional neural networks (CNNs). These methods can help us identify the defining characteristics of each phenotype. The proposed method for interpreting the unsupervised model will fall under the area of knowledge distillation and mimic learn-

ing [46]. Mimic learning is a means of interpreting or distilling knowledge captured by a less transparent model (called the 'teacher') by mimicking it through a more transparent model (called the 'student'). To the best of the author's knowledge, this work is the first attempt to interpret an unsupervised learning-based model through supervised learning. In particular, in our context, the 'teacher' model encapsulates the auto-encoder's encoder part, hierarchical clustering, followed by the assignment of phenotype labels. We propose to mimic this unsupervised pipeline through a 'student' CNN, for which we will formulate interpretability methods. CNNs extract features from input images, with the convolutional layers retaining the spatial information [24]. Several techniques have been developed to introduce class-specific interpretability within the CNN framework, with one notable method being the Gradient Class-Activation-Map (Grad-CAM) [45]. In this chapter, a volumetric extension of the Grad-CAM approach will be formulated that allows the production of interpretable visualizations for the discovered phenotypes at three levels of summarization: 3D, 2D, and 1D. These levels respectively indicate which voxels, which 2D spatial regions, and which slices depthwise are salient for each discovered phenotype.

The Grad-CAM method, as proposed by [45], is an effective CNN interpretability approach that has been widely adopted in various applications such as image classification, captioning, and visual question answering. The method operates on a pre-trained CNN and identifies the regions of feature maps from the last convolutional layer that are responsible for the network's final decision on each input. Specifically, the method calculates the gradient of the predicted class with respect to each feature map in the last convolutional layer, revealing the regions that significantly contribute to the classification output. In this work, a generalized volumetric extension of Grad-CAM is introduced that can handle higher dimensional tensors.

4.2 VolPAM, Volumetric Phenotype-Activation-Map

The CNN architecture used for the task of phenotype interpretability is a 3D CNN as presented in Figure 4.1. The input to this CNN was a 3D CT image volume, and it was trained to reproduce the phenotypes labels as assigned by the unsupervised approach of the last chapter. In a 3D CNN, the output of a convolutional layer is usually referred to as a feature-map tensor consisting of multiple 3D volumes called activation maps. As such, the feature-map tensor in itself is a 4D tensor of size $F \times H \times W \times D$, where F is the number of activation maps, and H, W, D represent



(a)

Layer (type)	Output Shape	Param #
Conv3d-1	[-1, 64, 128, 128, 64]	1,792
BatchNorm3d-2	[-1, 64, 128, 128, 64]	128
ReLU-3	[-1, 64, 128, 128, 64]	0
MaxPool3d-4	[-1, 64, 64, 64, 32]	0
Conv3d-5	[-1, 32, 64, 64, 32]	55,328
BatchNorm3d-6	[-1, 32, 64, 64, 32]	64
ReLU-7	[-1, 32, 64, 64, 32]	0
MaxPool3d-8	[-1, 32, 32, 32, 16]	0
Conv3d-9	[-1, 16, 32, 32, 16]	13,840
BatchNorm3d-10	[-1, 16, 32, 32, 16]	32
ReLU-11	[-1, 16, 32, 32, 16]	0
MaxPool3d-12	[-1, 16, 16, 16, 8]	0
Flatten-13	[-1, 32768]	0
Linear-14	[-1, 256]	8,388,864
ReLU-15	[-1, 256]	0
Linear-16	[-1, 4]	1,028

Total params: 8,461,076
Trainable params: 8,461,076
Non-trainable params: 0

(b)

Figure 4.1: CNN Model for Interpretability (a) Architecture: Classifier receives CT scan images (3D volumes) and is trained to produce, as output, their Phenotype labels as produced through Hierarchical Clustering. The input is passed through three convolutional layers with 64, 32, and 16 kernels, respectively. After that, the output of the last convolutional layer is flattened and passed through two linear (fully connected) layers with 256 and k^* (number of phenotypes, based on the Elbow method). Then, the classifier training process is finished, and the output of the last convolution layer will be used for computing Interpretable VolPAM maps. (b) Model summary: The interpretability CNN has been trained with a batch size of 10, and a learning rate of 0.001, using RMSProp optimizer, for 50 epochs. Different layers of the CNN, as well as their output shapes and the number of trainable parameters, are shown.

the height, width, and depth of the 3D feature map. The final layer of a classification neural network is a softmax with the same number of neurons as the number of classes, where the output of each neuron represents the probability assigned by the model for the input to have the corresponding class. Let c denote the class that is assigned the maximum probability by the model to the i-th example in the dataset, and let y_i^c denote that probability. The output of the last convolutional layer (shown as C3 in Figure 4.1) is what effectively is fed into the classification layer. CNNs preserve the spatial correspondence between the voxels of the input image volume and each 3D activation map from a convolutional layer. As a result, the focus will be on the activation maps of the C3-layer for interpreting the phenotype-labelling classifier. In particular, the interest here is to analyze how the output of the neural network i.e. the phenotype label, varies as a function of the voxels of activation maps in the C3-layer. This can be quantified by computing the gradient of the output with respect to each activation map. Let $\mathbf{A}_i^f$ denote the f-th feature map of the last convolutional layer for the i-th datapoint in the dataset. The gradient of y_i^c with respect to the f-th feature map is given by:

$$\mathcal{G}_i^{fc} = \frac{\partial y_i^c}{\partial \mathbf{A}_i^f} \tag{4.1}$$

The feature map, $\mathbf{A}_i^f$, is of size $H \times W \times D$, and y_i^c is a scalar. As such, $\mathcal{G}_i^{fc}$ is the derivative of a scalar with respect to a 3D volume and, therefore will also be of size $H \times W \times D$. The gradient $\mathcal{G}_i^{fc}$ is then averaged to compute its overall effect on the output for the i-th data point:

$$\alpha_i^{fc} = \frac{1}{HWD} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{d=1}^{D} \mathcal{G}_i^{fc}[h, w, d] \tag{4.2}$$

where $\mathcal{G}_i^{fc}[h, w, d]$ is a voxel of $\mathcal{G}_i^{fc}$ indexed by h, w, d.

Given that the phenotype is the output class in our scenario, we define the 3D Volumetric Phenotype Activation Map (VolPAM$_{3D}^c$) for a particular phenotype c as the weighted sum of activation maps, where the weights are determined by α_i^{fc}. This computation is carried out by averaging the weighted activation maps over all examples in the dataset that are classified by the CNN classifier as phenotype c.

$$VolPAM_{3D}^c = \frac{1}{N_c} \sum_i \delta(\hat{y}_i = c)(y_i^c) \left[ReLU \left(\sum_f \alpha_i^{fc} \mathbf{A}_i^{fc} \right) \right] \tag{4.3}$$

In the above expression, $\hat{y}_i$ denotes the phenotype that the CNN assigns to example i. The function $\delta(\hat{y}_i = c)$ takes the value of one if the assigned phenotype is c and zero otherwise. N_c represents the count of examples that are assigned phenotype c. It is important to note that we still incorporate the probability assigned by the model for an example to belong to phenotype c (y_i^c) to weight the individual contributions from the i-th example. The subscript $3D$ appended to $VolPAM_{3D}^c$ signifies that it is the voxel-level salience map for phenotype c.

We can create a 2D matrix to visualize the most relevant 2D spatial regions for phenotype c by computing the average of each pixel of $VolPAM_{3D}^c$ across the depth as follows:

$$VolPAM_{2D}^c = \frac{1}{D}\sum_{d=1}^{D} VolPAM_{3D}^c[:,:,d] \tag{4.4}$$

where $VolPAM_{2D}^c$ is $H \times W$, and the notation $VolPAM_{3D}^c[:,:,d]$ is used to represent a 2D slice from the volume $VolPAM_{3D}^c$ at depth index d.

To determine the significance of each slice in the volume, we calculate the average of all the pixels in $VolPAM_{3D}^c$ for each depth as follows:

$$VolPAM_{1D}^c = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} VolPAM_{3D}^c[h,w,:] \tag{4.5}$$

where $VolPAM_{1D}^c$ is $D \times 1$, and $VolPAM_{3D}^c[h,w,:]$ is a $D \times 1$ vector consisting of pixel (h,w) of $VolPAM_{3D}^c$ for all depths.

4.3 Results

In this section, the outcomes of experiments on the LIDC and COVID19-CT datasets are presented. As mentioned in the previous chapter, after doing representation learning through a convolutional autoencoder, we performed hierarchical clustering in the learned latent space (as shown in Figure 3.2). Using Equation 3.2, the optimal number of phenotypes (k^*) for the LIDC and COVID19-CT datasets were found to be two and three, respectively (Figures 3.3 and 3.4). Following the ideas introduced in this chapter, we proceeded to learn phenotype-label classifiers for each dataset with output classes equivalent to the k^* value corresponding to that dataset. We computed $2D$ and $1D$ salience maps for the discovered phenotypes for both datasets as shown in

Figures 4.2 and 4.3, respectively. The first row of each figure shows $VolPAM_{2D}$ (4.4), which represents the average of $3D$ attention heatmaps along the Z-axis, whereas the second row presents $VolPAM_{1D}$ (4.5), which represents the average of $3D$ attention heatmaps along the XY-plane. These $2D$ and $1D$ salience maps provide interpretable insights into the significance of the detected phenotypes.

Moreover, the $3D$ visualizations of $VolPAM_{3D}$ (4.3) have been uploaded as videos online (`https://tinyurl.com/VolPAM3DVisualizations`), enabling clinicians and radiologists to interact with the data and gain further insights into the detected phenotypes.

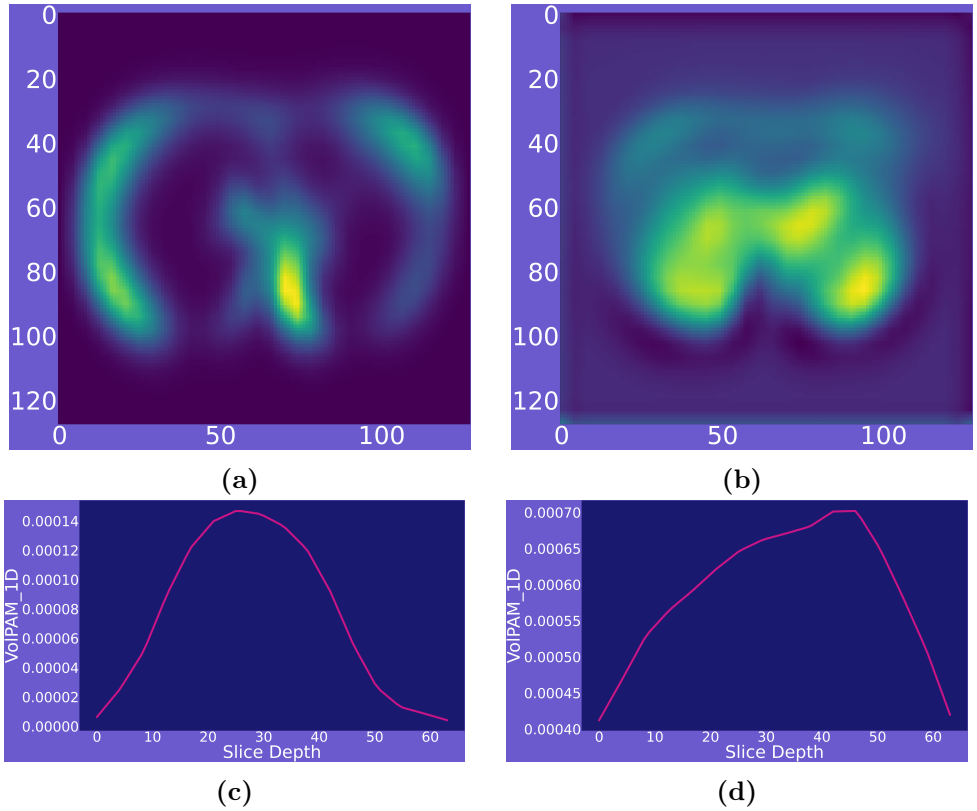


Figure 4.2: The first row of the visualizations shows the average $VolPAM_{3D}$ heatmaps over the Z-axis for all CT scans that belong to either $Cluster0$ or $Cluster1$ in the LIDC dataset. The second row displays the average $VolPAM_{2D}$ maps over the XY-axis for the same clusters, which are obtained by averaging the $VolPAM_{3D}$ heatmaps.

The results reveal several noteworthy observations. Firstly, upon further analysis of the LIDC dataset (Figure 4.2), the first phenotype (Fig 4.2a) demonstrated higher attention values for one of the lungs in an overall 2D sense averaged across all depths, while the second phenotype (Fig 4.2b) showed hotspots in both lungs. In terms

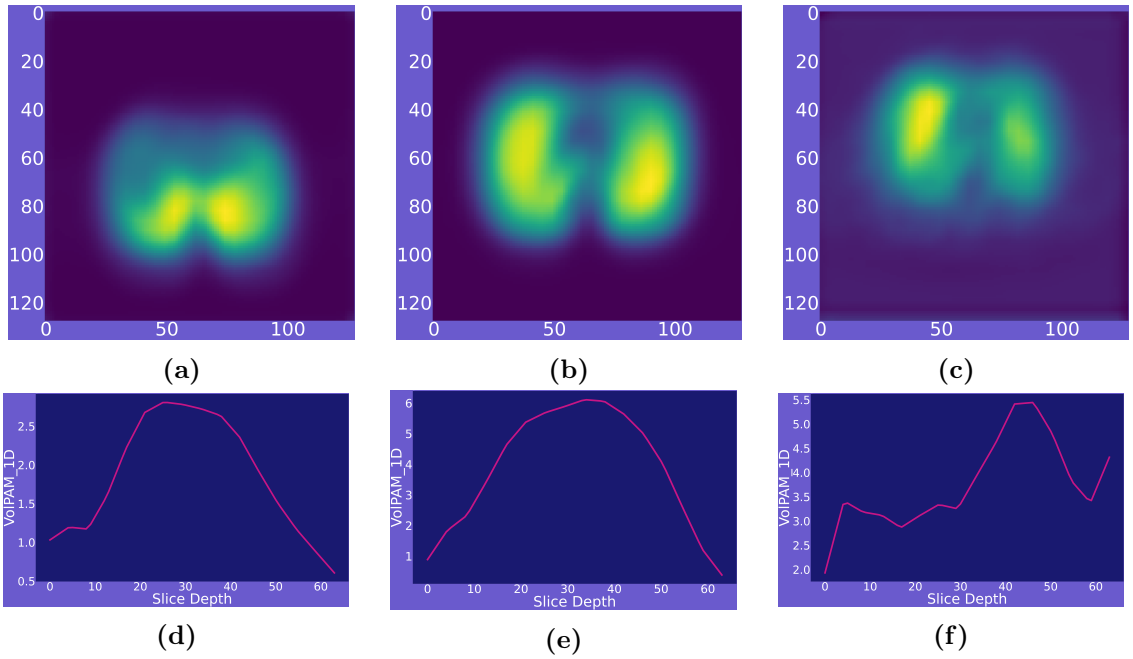


Figure 4.3: The visualizations for different clusters in the Covid19-CT dataset include $VolPAM_{2D}$ and $VolPAM_{1D}$. The first row of the plots shows the average of $VolPAM_{3D}$ heatmaps over the Z-axis for all CT scans in $Cluster0$, $Cluster1$, and $Cluster2$. The second row of the plots shows the average of $VolPAM_{2D}$ maps (based on the first row pictures) over the XY-axis for all the CT scans that belong to $Cluster0$, $Cluster1$, and $Cluster2$, respectively.

of depth, the first phenotype (Fig 4.2c) indicated higher salience values for middle slices, while the second phenotype paid more attention to deeper slices in the volume. Both phenotypes showed little or no attention to slices at the extreme ends of the volume. This demonstrates that the voxels from slices corresponding to extreme ends of the volume are not contributing toward phenotype-label classification. Similarly, for the COVID dataset, the $2D$ attention maps (Figures 4.3 a, b, c) demonstrated different lung areas as regions of interest for each phenotype. Specifically, regarding the depthwise salience of phenotypes (Figures 4.3 d, e, f), the first two phenotypes focused on middle slices, while the third phenotype emphasized slices deeper along the Z-axis, but still the values taper off near the extreme end of the volume. Additionally, these findings also provide a sanity check for the discovered phenotypes in that the unsupervised approach of the previous chapter was not overfitting to irrelevant areas of the imaging volume.

4.4 Summary

In this chapter, a data-driven approach is proposed for interpreting the phenotypes discovered through unsupervised representation learning and hierarchical clustering. The method introduces supervised mimic learning, where a neural network is trained to mimic the unsupervised pipeline. The supervised model learns to classify the phenotype labels assigned by the unsupervised approach. By leveraging interpretability methods developed for supervised neural networks, the phenotypes' defining characteristics are identified. A volumetric extension of the Gradient Class-Activation-Map (Grad-CAM) approach is introduced to produce interpretable visualizations at different levels of summarization: 3D, 2D, and 1D. The method is applied to the LIDC and COVID19-CT datasets, and the results are presented in terms of 2D and 1D salience maps. The visualizations provide insights into the significance of the detected phenotypes and validate the relevance of the unsupervised approach.

Our experiments have been performed using NVIDIA RTX A4000 GPU. Using this server as our resource, this part of our work took 2 to 3 hours of processing for each dataset. However, producing the 3D, 2D, and 1D VolPAMs for one CT scan study only takes a few minutes, which is promising.

Chapter 5

Conclusions and Future Directions

5.1 Conclusions

In this thesis, an unsupervised pipeline for representation learning and phenotype discovery from 3D imaging volumes has been presented. A novel approach, VolPAM, has been introduced to interpret the discovered phenotypes by mimicking the unsupervised pipeline through a deep neural network and providing interpretability of the learned model. The potential of this approach is to complement the work of clinicians and radiologists by offering insight into potential subtype profiles and their significance in new or partially understood diseases.

The use of convolutional neural networks (CNNs) for image processing and computer vision tasks, particularly in medical imaging, has been explored. CNNs have demonstrated remarkable performance in various computer vision applications, including object recognition and image classification. Important works in the field have been reviewed, emphasizing the breakthrough performance of CNNs in the 2012 ILSVRC object recognition competition.

In conclusion, a framework that combines unsupervised and supervised learning paradigms for phenotype discovery and interpretability has been presented. The methods introduced in this thesis have been evaluated on lung CT datasets, demonstrating their potential for analyzing 3D imaging volumes. Further research is necessary to investigate the feasibility and effectiveness of this approach on other imaging modalities, such as multimodality MRI.

5.2 Future Directions

In this section, potential future research directions based on the findings and limitations of the present study are discussed.

Firstly, the accuracy and generalizability of lung area segmentation in the VolPAM pipeline can be improved by exploring deep learning-based segmentation methods, such as 3D UNet. Much improved results have been shown by these methods in various applications, allowing them to capture complex and robust features from large-scale data.

Secondly, deep learning approaches can enhance the estimation of the optimal number of clusters in unsupervised clustering. By combining deep representation learning with a neural network-based deep cluster estimation module, it can be possible to obtain the optimal number of clusters and cluster labels for the data without relying on distance metrics or heuristics.

Exploring other autoencoder architectures, such as sequence-to-sequence autoencoders based on recurrent or transformer networks, can further improve representation learning. These architectures are capable of handling sequential data and capturing long-term dependencies and sequential information, which are valuable in the context of imaging volumes.

Utilizing clustering validation techniques such as the Silhouette and Rand index to quantify and validate clustering results. These metrics play a vital role in validating clustering outcomes and selecting the best cluster count.

The integration of attention maps generated by VolPAM with domain-specific, biomedical, and multimodal large language models can provide easy-to-understand textual interpretation of learned representations. This integration facilitates the communication of discovered phenotypes and their clinical significance to healthcare professionals.

Finally, the development of an interactive visualization tool for VolPAM maps enables clinicians and radiologists to manipulate and explore the salience maps in 3D, 2D, and 1D. This tool provides functionalities such as zooming, rotating, slicing, measuring, and annotating, thereby enhancing the interpretation and analysis of the discovered phenotypes.

These future research directions hold the potential to advance the field of unsupervised representation learning and phenotype discovery from 3D imaging volumes. Such methods are expected to contribute to the development of interpretable and

clinically relevant AI models in medical imaging.

Bibliography

[1] Emmanuel Rios Velazquez, Chintan Parmar, Ying Liu, Thibaud P. Coroller, Gisele Cruz, Olya Stringfield, Zhaoxiang Ye, Mike Makrigiorgos, Fiona Fennessy, Raymond H. Mak, Robert Gillies, John Quackenbush, and Hugo J.W.L. Aerts. Somatic mutations drive distinct imaging phenotypes in lung cancer. *Cancer Research*, 77(14):3922–3930, 2017.

[2] Cho N. Molecular subtypes and imaging phenotypes of breast cancer. *Ultrasonography*, 35(4):281–288, 2016.

[3] J.B. Savitz and W.C. Drevets. Imaging phenotypes of major depressive disorder: genetic correlates. *Neuroscience*, 164(1):300–330, 2009.

[4] Jacob W. Vogel, Alexandra L. Young, Neil P. Oxtoby, Ruben Smith, Rik Ossenkoppele, Leon M. Aksman, Olof Strandberg, Renaud La Joie, Michel Grothe, Yasser Iturria Medina, Gil D. Rabinovici, Daniel C. Alexander, Alan C. Evans, and Oskar Hansson. Spatiotemporal imaging phenotypes of tau pathology in alzheimer's disease. *Alzheimer's & Dementia*, 16(S4):e045612, 2020.

[5] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

[6] S. Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S. Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L. Prince, Daniel Rueckert, and Ronald M. Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.

[7] Ravi Aggarwal, Viknesh Sounderajah, Guy Martin, Daniel S W Ting, Alan Karthikesalingam, Dominic King, Hutan Ashrafian, and Ara Darzi. Diagnostic

accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digital Medicine*, 4(1):65, 2021.

[8] Christoph Sommer, Rudolf Hoefler, Matthias Samwer, and Daniel W. Gerlich. A deep learning and novelty detection framework for rapid phenotyping in high-content screening. *Molecular Biology of the Cell*, 28(23):3428–3436, 2017.

[9] Oliver Dürr, Elvis Murina, Daniel Siegismund, Vasily Tolkachev, Stephan Steigele, and Beate Sick. Know when you don't know: A robust deep learning approach in the presence of unknown phenotypes. *ASSAY and Drug Development Technologies*, 16(6):343–349, 2018.

[10] Ashraf AB, Daye D, Gavenonis S, Mies C, Feldman M, Rosen M, and Kontos D. Identification of intrinsic imaging phenotypes for breast cancer tumors: preliminary associations with gene expression profiles. *Radiology*, 272(2):374–384, 2014.

[11] Despina Kontos, Stacey J. Winham, Andrew Oustimov, Lauren Pantalone, Meng-Kang Hsieh, Aimilia Gastounioti, Dana H. Whaley, Carrie B. Hruska, Karla Kerlikowske, Kathleen Brandt, Emily F. Conant, and Celine M. Vachon. Radiomic phenotypes of mammographic parenchymal complexity: Toward augmenting breast density in breast cancer risk assessment. *Radiology*, 290(1):41–49, 2019.

[12] M. Norouzi, S. Khan, and A. Ashraf. Volpam: Volumetric phenotype-activation-maps for data-driven discovery of 3d imaging phenotypes and interpretability. *Neural Computing and Applications*, 2023 (Under Review).

[13] Caiming Zhang and Yang Lu. Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration*, 23:100224, 2021.

[14] Damini Dey, Piotr J Slomka, Paul Leeson, Dorin Comaniciu, Sirish Shrestha, Partho P Sengupta, and Thomas H Marwick. Artificial intelligence in cardiovascular imaging: Jacc state-of-the-art review. *Journal of the American College of Cardiology*, 73(11):1317–1335, 2019.

[15] Stuart Rusell and Peter Norvig. Artificial intelligence: A modern approach. *Pretice Hall Series in Artificial Intelligence*, 1:649–789, 2003.

[16] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[17] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR 2018*, 2018.

[18] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2547–2555, 2019.

[19] Min Chen, Xiaobo Shi, Yin Zhang, Di Wu, and Mohsen Guizani. Deep feature learning for medical image analysis with convolutional autoencoder neural network. *IEEE Transactions on Big Data*, 7(4):750–758, 2017.

[20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[22] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[23] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[24] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9(4):611–629, 2018.

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[27] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 2021.

[28] Sakshi Patel, Shivani Sihmar, and Aman Jatain. A study of hierarchical clustering algorithms. In *2015 2nd international conference on computing for sustainable global development (INDIACom)*, pages 537–541. IEEE, 2015.

[29] Xifeng Guo, Xinwang Liu, En Zhu, and Jianping Yin. Deep clustering with convolutional autoencoders. In *International conference on neural information processing*, pages 373–382. Springer, 2017.

[30] Shehroz S Khan and Amir Ahmad. Cluster center initialization algorithm for k-means clustering. *Pattern recognition letters*, 25(11):1293–1302, 2004.

[31] Ying Zhao, George Karypis, and Usama Fayyad. Hierarchical clustering algorithms for document datasets. *Data mining and knowledge discovery*, 10(2):141–168, 2005.

[32] Anil K Jain and Richard C Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[33] Yogita Rani[1] and Harish Rohil. A study of hierarchical clustering algorithm. *ter S & on Te SIT*, 2:113, 2013.

[34] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[35] Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *Advances in neural information processing systems*, 6, 1993.

[36] Nathalie Japkowicz, Stephen Jose Hanson, and Mark A Gluck. Nonlinear autoassociation is not equivalent to pca. *Neural computation*, 12(3):531–545, 2000.

[37] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[38] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49. JMLR Workshop and Conference Proceedings, 2012.

[39] Yifei Zhang. A better autoencoder for image: Convolutional autoencoder. In *ICONIP17-DCEC. Available online: http://users. cecs. anu. edu. au/Tom. Gedeon/conf/ABCs2018/paper/ABCs2018_ paper_ 58. pdf (accessed on 23 March 2017)*, 2018.

[40] Ali Alqahtani, Xianghua Xie, Jingjing Deng, and Mark W Jones. A deep convolutional auto-encoder with embedded clustering. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 4058–4062. IEEE, 2018.

[41] Yueqing Wang, Zhige Xie, Kai Xu, Yong Dou, and Yuanwu Lei. An efficient and effective convolutional auto-encoder extreme learning machine network for 3d feature learning. *Neurocomputing*, 174:988–998, 2016.

[42] L. Yan and etal. An interpretable mortality prediction model for covid-19 patients. *Nature Mach. Intell*, 2(5):283–288, 2020.

[43] Aditya Saini and Ranjitha Prasad. Select wisely and explain: Active learning and probabilistic local post-hoc explainability. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 599–608, 2022.

[44] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[45] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[46] Guo-Hua Wang, Yifan Ge, and Jianxin Wu. Distilling knowledge by mimicking features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8183–8195, 2022.

[47] OECD Paris. Oecd publishing; 2017. health at a glance 2017: Oecd indicators.

[48] Geoffrey D Rubin, Christopher J Ryerson, Linda B Haramati, Nicola Sverzellati, Jeffrey P Kanne, Suhail Raoof, Neil W Schluger, Annalisa Volpi, Jae-Joon Yim,

Ian BK Martin, et al. The role of chest imaging in patient management during the covid-19 pandemic: a multinational consensus statement from the fleischner society. *Radiology*, 296(1):172–180, 2020.

[49] Ho Yuen Frank Wong, Hiu Yin Sonia Lam, Ambrose Ho-Tung Fong, Siu Ting Leung, Thomas Wing-Yan Chin, Christine Shing Yen Lo, Macy Mei-Sze Lui, Jonan Chun Yin Lee, Keith Wan-Hang Chiu, Tom Wai-Hin Chung, et al. Frequency and distribution of chest radiographic findings in patients positive for covid-19. *Radiology*, 296(2):E72–E78, 2020.

[50] Cheng Jin, Weixiang Chen, Yukun Cao, Zhanwei Xu, Zimeng Tan, Xin Zhang, Lei Deng, Chuansheng Zheng, Jie Zhou, Heshui Shi, et al. Development and evaluation of an artificial intelligence system for covid-19 diagnosis. *Nature communications*, 11(1):1–14, 2020.

[51] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510, 2018.

[52] Peter Szolovits, Ramesh S Patil, and William B Schwartz. Artificial intelligence in medical diagnosis. *Annals of internal medicine*, 108(1):80–87, 1988.

[53] Alison M Darcy, Alan K Louie, and Laura Weiss Roberts. Machine learning and the profession of medicine. *Jama*, 315(6):551–552, 2016.

[54] Xiaoli Tang. The role of artificial intelligence in medical imaging research. *BJR| Open*, 2(1):20190031, 2019.

[55] Morgan P McBee, Omer A Awan, Andrew T Colucci, Comeron W Ghobadi, Nadja Kadom, Akash P Kansagra, Srini Tridandapani, and William F Auffermann. Deep learning in radiology. *Academic radiology*, 25(11):1472–1480, 2018.

[56] Curtis P Langlotz, Bibb Allen, Bradley J Erickson, Jayashree Kalpathy-Cramer, Keith Bigelow, Tessa S Cook, Adam E Flanders, Matthew P Lungren, David S Mendelson, Jeffrey D Rudie, et al. A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 nih/rsna/acr/the academy workshop. *Radiology*, 291(3):781, 2019.

[57] Lei Cai, Jingyang Gao, and Di Zhao. A review of the application of deep learning in medical image classification and segmentation. *Annals of translational medicine*, 8(11), 2020.

[58] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[59] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[60] Min Chen, Yujun Ma, Yong Li, Di Wu, Yin Zhang, and Chan-Hyun Youn. Wearable 2.0: Enabling human-cloud integration in next generation healthcare systems. *IEEE Communications Magazine*, 55(1):54–61, 2017.

[61] David C Levin, Laurence Parker, and Vijay M Rao. Recent trends in imaging use in hospital settings: implications for future planning. *Journal of the American College of Radiology*, 14(3):331–336, 2017.

[62] Ahmed Bilal Ashraf, Dania Daye, Sara Gavenonis, Carolyn Mies, Michael Feldman, Mark Rosen, and Despina Kontos. Identification of intrinsic imaging phenotypes for breast cancer tumors: preliminary associations with gene expression profiles. *Radiology*, 272(2):374, 2014.

[63] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

[64] Shokouh Shakouri, Mohammad Amin Bakhshali, Parvaneh Layegh, Behzad Kiani, Farid Masoumi, Saeedeh Ataei Nakhaei, and Sayyed Mostafa Mostafavi. Covid19-ct-dataset: an open-access chest ct image repository of 1000+ patients with confirmed covid-19 diagnosis. *BMC Research Notes*, 14(1):1–3, 2021.

[65] Yi-Chen Chen, Sung-Feng Huang, Hung-yi Lee, Yu-Hsuan Wang, and Chia-Hao Shen. Audio word2vec: Sequence-to-sequence autoencoding for unsupervised

learning of audio segmentation and representation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(9):1481–1493, 2019.

[66] Anega Maheshwari, Priyanka Mitra, and Bhavna Sharma. Autoencoder: Issues, challenges and future prospect. In *Recent Innovations in Mechanical Engineering: Select Proceedings of ICRITDME 2020*, pages 257–266. Springer, 2022.

[67] Qingkai Kong, Andrea Chiang, Ana C Aguiar, M Giselle Fernández-Godino, Stephen C Myers, and Donald D Lucas. Deep convolutional autoencoders as generic feature extractors in seismological applications. *Artificial Intelligence in Geosciences*, 2:96–106, 2021.

[68] K Sasirekha and P Baby. Agglomerative hierarchical clustering algorithm-a. *International Journal of Scientific and Research Publications*, 83(3):83, 2013.

[69] MS Yang. A survey of hierarchical clustering. *Mathl. Comput. Modelling*, 18(11):1–16, 1993.

[70] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.

[71] Himanshu Mittal, Avinash Chandra Pandey, Mukesh Saraswat, Sumit Kumar, Raju Pal, and Garv Modwel. A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets. *Multimedia Tools and Applications*, pages 1–26, 2021.

[72] MA Syakur, BK Khotimah, EMS Rochman, and Budi Dwi Satoto. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering*, volume 336, page 012017. IOP Publishing, 2018.