People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

University M'Hamed BOUGARA – Boumerdès

Institute of Electrical and Electronic Engineering

Department of Power and Control

Project Report Presented in Partial Fulfilment of the Requirements of the Degree of

MASTER

In **Electrical and Electronics Engineering**

Option**: Control Engineering**

Title:

Hierarchical Clustering Technique for Lung Diseases

Presented by:

- **Inas KOUFI**

Supervisor:

Prof. Dr. Abdelmalek KOUADRI

Registration number: ……/2022

Abstract

Each year, pulmonary diseases are the underlying cause of death worldwide. The process of detecting lung diseases can be time-consuming and error-prone. Such errors can be expensive and affect patients' lives. Accuracy and fast diagnosis are therefore crucial. Due to its high clinical impact and remaining challenges, medical image analysis has become a broad and active area of research in recent decades where various machine learning methods have been developed to assist in medical diagnosis.

These machine learning models often use neural networks as a tool of image manipulation, feature extraction, and classification and clustering techniques. Our proposed solution consists of using an alternative approach. In our study, distance and similarity measures have been applied to medical images hierarchical clustering in order to determine their usability and accuracy in detecting lung diseases. Three indices were covered in this work in order to cluster chest X-rays: Euclidean distance, cosine distance, and Jensen-Shannon divergence.

Each metric proposed has been applied to and tested on CheXphoto dataset with seven labels. Promising results were obtained with an accuracy range of 61.6% to 81.2% of correct predictions. Therefore, the proposed methods have a good application prospect and promotion value.

Keywords: Image Similarity, Machine Learning (ML), Hierarchical Clustering, Jensen-Shannon Divergence (JSD), Euclidean Distance, Cosine Distance, CheXphoto dataset.

Dedication

To my parents and siblings
for being a source of motivation and inspiration.

To my past, present, and future friends
for being companions along a rough path.

To the Blues
for fueling me with passion when there was nowhere to stem from.

To those who have always believed I could do great things despite life.

Acknowledgment

First and foremost, all praise to Allah, the most gracious, the most merciful, for providing me strength, courage, and patience to complete this humble work.

Secondly, I express my dearest gratitude and my sincerest respect to my supervisor, Prof. A. Kouadri, for the continuous guidance, help, and most importantly for believing in me when I did not believe in myself. It is and forever will be my greatest honor to have completed this work under his supervision.

Finally, to my teachers, faculty staff, colleagues, and mentors who have contributed directly or not to the success of this thesis.

Table of Contents

List of Figures

List of Tables

Nomenclature

ACC	Accuracy
AI	Artificial Intelligence
CAD	Computer-Aided Systems
DL	Deep Learning
GPU	Graphical Processing Unit
JSD	Jensen-Shannon Divergence
KLD	Kullback-Leibler Divergence
LR	Likelihood Ratio
ML	Machine Learning
NN	Neural Network
RGB	Red Green Blue
$SOTA$	State-of-the-art

General Introduction

Respiratory diseases impose a global health burden for being a leading cause of morbidity, disability, and mortality worldwide. Lungs are constantly exposed to a myriad of noxious agents present in ambient air, such as particles, chemicals and infectious organisms. Early detection of lung diseases plays a major role in the chance of disease recovery and treatment. Timely detection of respiratory diseases requires adequate access to, and use of diagnostic instruments, in addition to effective and practical case-finding approaches that requires intervention of medical doctors or healthcare professionals.

This is becoming a very complex task if a radiologist is not available at time. For this reason, researchers and scientists have been focusing on creating computer aided systems that assist medical doctors in the detection of lung diseases. These systems use artificial intelligence to facilitate the detection of pulmonary diseases using chest imaging. As a diagnostic tool, medical imaging is one of the most revolutionary advances in medicine. By providing a visual representation of the inside of human body, medical imaging helps radiologists make earlier and more accurate diagnosis. Thus, diseases can be treated more effectively to improve the quality of patient care.

Artificial intelligence and automatic image analysis allow a significant reduction in work load and a swifter early detection of diseases. These methods often combine handcrafted feature representations and classifiers. Many of the previous methods and algorithms have been proposed across the past decades to allow such assistant in diagnosing lung diseases. The purpose of these algorithms is to automatically uncover patterns in data. One way to realize this, is by using hierarchical clustering. However, they required both large volume of dataset and strong computational power computers.

The approach proposed in this thesis leverages existing tool - instead of creating new AI algorithms - to test their accuracy in medical diagnosis. The motivation behind this project was to create a pipeline that allows the use of small datasets ($< 10^3$) and limited computational power. Distance and similarity measures are presented as a tool focusing our algorithm on as a tool of diagnosis. Such methods have been applied to several AI applications and have performed relatively well. This thesis, however, focuses on the gap existing when implementing such metrics in medical image diagnosis. Three methods have been proposed: Euclidean distance, Cosine distance, and Jensen-Shannon divergence. Each method has been used to cluster images, then tested for accuracy and purity of clustering.

The rest of this thesis is structured as follows. Chapter 1 first introduces the motivation behind this study, a strong theoretical background relevant to the scope of this project, and reviewing the latest related research done to detect lung diseases from X-ray

images. Chapter 2 presents the tools used in this work, as well as a detailed overview of the proposed methods and their evaluation techniques. Chapter 3 covers the application, experimental results, and the effectiveness of each proposed approach in clustering lung diseases. Finally, this thesis is concluded by discussing findings and future work directions.

Chapter 1

Pulmonary Diseases and Machine Learning

1.1 Introduction

Pulmonary diseases, or diseases of the lung, represent some of the most propagated and fatal conditions across the globe. Timely detection plays a major role in the advancement of treatment and recovery. Many tools have been explored over the past decades to insure that. However, the latest statistics prove that the efforts endured are still not enough to prevent the increasing number of mortal cases.

In this chapter, relevant technologies, theories, and proposed approaches adopted in detection of lung diseases are explained. An overview on medical image analysis is given, while identifying state-of-the-art (SOTA) approaches for disease detection in chest X-rays. The objective of this chapter is to first, shed light on the importance of this research scope. Second, is to present current research efforts and what is known about the methods that have already been explored.Finally, to indicate the knowledge gaps in the research area, and the challenges that are faced.

1.2 Pulmonary Diseases Overview

Lung diseases are among the leading causes of death worldwide. A study conducted in 2017 shows that 544.9 million people have a chronic respiratory disease, representing an increase of 39.8% compared with 1990 [1]. The European Lung White Book argues that lung infection, lung cancer, and chronic obstructive pulmonary disease (COPD) together are accounted for 9.5 million deaths during 2008, one-sixth of the global total [2]. Moreover, 1.76 million people die from lung cancer each year, making it the deadliest cancer in the world [3]. Pulmonary diseases do not only affect adults. Pneumonia kills over 800,000 children under 5 years old yearly; one child every 39 seconds [4].

In Algeria, statistics revealed that lung diseases have a large rate of mortality among the population. Influenza and pneumonia are ranked 7^{th} leading causes of death, followed by lung cancer in the 9^{th} place, and tuberculosis in the 14^{th} place [5]. Lung cancer is the third most common cancer in the country; broken-down by gender, it is the most common within men. Lung cancer cases have reached an increase of around 8.2% every year. In regards of fatality, lung cancer has over 13.2% mortal cases, making it the deadliest cancer in the country [6].

Economic transition, lifestyle changes, and pollution have increased the incidence of these diseases across the world. The high prevalence of respiratory infections, household, and outdoor pollution, increased tobacco epidemic, and lifestyle changes that come with urbanization all together have made respiratory diseases a major burden. The large numbers presented above are caused by various factors. A World Health Organization (WHO) report claims that 4.2 million people die every year as a result of exposure to ambient air pollution. Whereas 3.8 million people die every year as a result of household exposure to smoke from dirty cook-stoves and fuels. These risks are increased given that 91% of the world's population live in places where air quality exceeds WHO guideline limits.[7]. Furthermore, the lack of hospitals, trained professionals, and medical diagnostic equipment are other challenges that make the diagnosis and treatment of these diseases difficult.

1.3 Pulmonary Disease Detection

In their primitive stages, lung diseases are detected with the aid of blood test, skin test, and some X-ray and CT scan. Chest X-rays are the most cost-effective and widely used tools used for examining and finding the abnormalities and specific structures present in lungs [8] [9]. Nowadays, with technological advancements, machine learning algorithms are used to investigate patterns in medical images that results in disease detection. A prior understanding of X-ray imaging and chest X-rays is required.

1.3.1 Medical Imaging (Radiography)

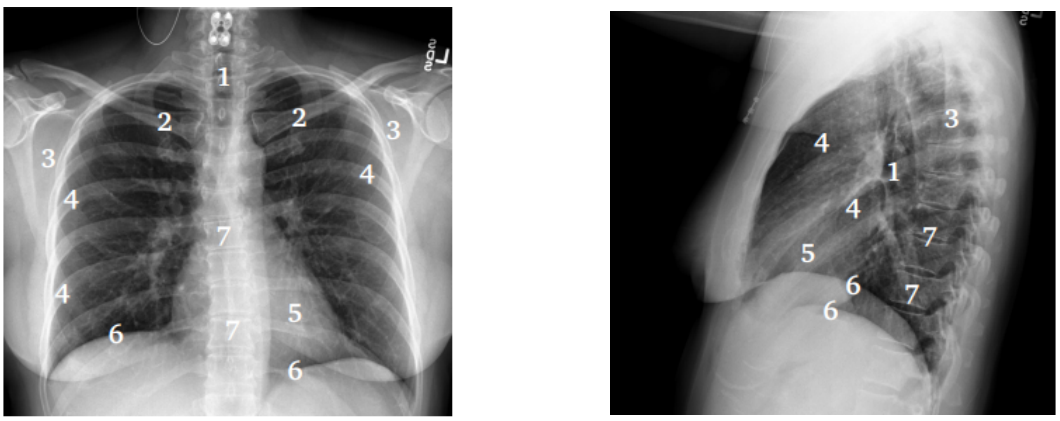
Medical imaging techniques can provide detailed images of human anatomy by creating visual representations of the body's interior in order to make an accurate diagnosis. Medical imaging techniques assist physicians by providing additional information that cannot be seen by a physical examination. Hence, helping physicians avoid medical mistakes. These images reveal details about structural, functional organs, and tissues.

Radiography (or X-ray imaging) is the most common imaging technique due to its accessibility and affordability for the assessment of lung diseases. In the interpretation of x-rays, one must consider different factors affecting the quality of X-ray images. One is a technical factor that includes exposure to film (patient radiation exposure), positioning, penetration, inspirational effort volumes, and the availability of artifacts. A quality X-ray is characterized by a posterior-anterior (PA) view, a full inspiratory effort where 10 posterior ribs are visible and the clavicle bones seen equidistant from the vertebral spinous processes. Some of the factors are summarized in the following table [10],[11]:

Factors	Analysis
High Quality	Well penetrated, vertebral bodies visible behind the heart
Low Quality	Under penetrated film-whitened and vertebral bodies not visible.
Frontal View	Contains much information on the thoracic cage, pleura, lungs, pericardium, heart, mediastinum and upper abdomen.
Lateral View	Usually taken to complement the frontal view.
PA Positioning	Posterior Anterior: Heart and mediastinum are closest to the film and therefore not magnified (Commonly used)
AP Positioning	Anterior Posterior: Used for patients who are on a chair or in bed. The clavicle bones cover the upper parts of the lung. The size of the heart and the mediastinum are also magnified.
Inspiratory efforts	10 ribs should be visible, whereas in poor inspiratory efforts, less than 6 ribs are visible.
Artifacts	Clearly seen on x-ray with well-defined opacities. Artifacts such as patients with tubes, pacemakers, and clothing (buttons).
Male & Female chest	The amount of breast tissue creates a difference breast tissue absorbs some of the x-rays may cause under or overexposure bilateral basilar lung infiltrates may cause asymmetrical lung density.
Lungs	Normal air-filled lungs appear black, each zone must be compared in the right and the left.
Mediastinum	Heart- Cardiothoracic Ratio-The maximum transverse diameter of the heart should not exceed 50% of the maximum transverse diameter of the chest on a standard posteroanterior (PA) radio-graph
Bones	The clavicles, ribs, and the spine .

Table 1.1: Chest X-ray imaging factors.

The following image presents a clearer illustration of the features of a chest X-ray.



(a) Frontal.

(b) Lateral.

Figure 1.1: Typical examination type in which two corresponding X-ray images of the chest are taken from one patient. On both, the following anatomical structures are visible: (1) trachea, (2) clavicle, (3) scapulae, (4) ribs, (5) heart, (6) diaphragm, and (7) vertebrae forming the spine. [12]

The interpretation of X-rays is a task that requires expertise and technical knowledge of human anatomy and pathologies. Hence, X-rays need to be examined by radiologists or healthcare professionals who have enough knowledge and experience to determine the proper diagnosis. This task often becomes laborious in the absence of a healthcare professional. With the technological developments reached nowadays, pulmonary disease detection is eased thanks to computer-aided systems and artificial intelligence.

1.3.2 Artificial Intelligence in Medical Diagnosis

The task of lung disease detection is to uncover physical signs when a patient poses challenging problems by combining medical knowledge, judgment, and experience usually with the help of computer-based systems. These systems are commonly known as computer-aided diagnostic (CAD) systems.

According to [13], CAD systems were first introduced in the late 1950s where biomedical scientists started investigating the use of a computer in solving medical and biological problems. Afterwards, CAD systems began to serve as a foundation for accelerating research in the interface of medicine and computer science. Its applications in searching for abnormalities uses medical imaging techniques [14].

CAD systems often use artificial intelligence (AI) to serve as diagnostic systems. Researches have built and developed machine learning (ML) algorithms to deal with this matter. Therefore, the role played by AI algorithms in facilitating detection of lung diseases from X-rays is crucial and significant. The major downside of ML approaches is that they principally rely on handcrafted features (manual engineering) that involve time-consuming and tedious tasks (labor-intensive tasks)[15].

Deep learning (DL), a sub-field of ML, appears to overcome this challenging issue. Difficulties arise, however, because of the data-hunger characteristics of the DL algorithms in which the availability of such huge datasets are quite limited. Since the vast majority of the existing studies are conducted on historical datasets to train algorithms, they showed performance drops when encountered in real-case scenarios [16].
Authors in [17] and [18] have employed DL-based techniques that have achieved some improvements. However, some conditions were still challenging to detect. Pneumonia, nodule, and infiltration achieved less that 70% detection rate with an AUC (area under the curve) of 0.6902. Even though major efforts have been done and have resulted in tremendous achievements, there still are some questions unanswered. Considering the latest major studies [17], [18], [19], [20], [21], [22], [23], and [24], the AUC measured is found to be in the range between 0.7471 and 0.808 only, indicating around 20% misclassification errors. Therefore, the results are somehow unconvincing.

The training of a DL model, usually a neural network (NN), for lung disease classification arises a challenge: the mismatch between the small input size of the neural network and the large image size of chest X-rays, and the wide variety of diagnoses. Modern chest X-rays today typically have a size of 2000 pixels to 3000 pixels [25] due to their high spatial resolution since high spatial resolution is required by radiologists to distinguish the small details of various lung pathologies. Moreover, the input size of common convolutional neural networks (CNN) for image classification in computer vision is ap-

proximately 224 pixels to 299 pixels [26]. Hence, to correct this discrepancy, the original image is often downsized to match the CNN input size. This process reduces the spatial resolution and can compromise the visibility of important features in the image. Another major issue with DL methods is that they require large computational powers. This task can be wearisome given that hospitals could find this solution useful, especially due to data confidentiality and graphical processing unit (GPU) accessibility.

1.4 Machine Learning

1.4.1 Definition

Machine Learning (ML) is a sub-field of AI. It is the area of study where computer programs learn from experience with respect to some tasks and performance measure. ML allows these algorithms to improve with experience and become more accurate in regards of making predictions. There are four main sub-classes of ML, categorized based upon how an algorithm learns; supervised, unsupervised, semi-supervised, and reinforcement learning.

Supervised Learning

The dataset used in this type of ML is labeled, and the variables to be assessed for correlation are defined. Hence, both the input and output of the algorithm are specified. These algorithms are best for:

- Binary classification: dividing data into two categories.
- Multi-class classification: choosing between more than two classes.
- Regression problem: predicting continuous values.
- Ensembling: combining predictions of multiple ML models to produce an accurate prediction.

Unsupervised Learning

This type involves unlabeled dataset. The algorithm scans through the training dataset looking for patterns and connections. The training dataset and the predictions are both predetermined. Unsupervised algorithms are good for the following tasks:

- Clustering: splitting dataset into groups based on similarity.
- Anomaly detection: identifying unusual data points in the dataset.
- Association mining: identifying sets of items in the dataset that frequently occur together.
- Dimentionality reduction: reducing the number of variables in the dataset.

Semi-supervised Learning

This approach involves a mixture of the two previous ones. The algorithm is fed with labeled training set, however, it is also free to explore the data on its own to develop its own understanding of the dataset. Some areas that use semi-supervised learning include:

- Machine translation: translation algorithms based on less than a full dictionary of words.
- Fraud detection: identifying fraud detection given only few example cases.
- Labeling data: algorithms trained on small datasets to apply data labels on a large set automatically.

Reinforcement Learning

It works by building a model with a distinct goal and a prescribed set of rules for accomplishing the goal. It is often used in:

- Video game-play: teaching bots to play a number of video games.
- Resources management: helping enterprise plan out how to allocate finite resources to reach a predefined goal.

1.4.2 The Learning Process

To construct a model, it must go through a learning process. The latter consists of these steps:

Data Acquisition

Data is obtained from the real physical world through measuring devices that convert physical quantities and observation into digital data that can be read and processed by computers.

Preprocessing

Preprocessing refers to the transformation of input data befor exploiting it by the algorithm. Preprocessing can be in the form of:

- **Data Cleaning**:
  Data acquired from the real world tends to be noisy, incomplete, and inconsistent. Data cleaning refers to the process of smoothing noise, filling in missing data, and correct inconsistencies.
- **Standarization**:
  A transformation of attributes with a Gaussian distribution and different means and standard deviations to a standard Gaussian distribution of mean zero and standard deviation of 1. This is realized by the following transformation: $z = \frac{x-\mu}{\sigma}$, where μ is the mean, σ is the standard deviation, x is the original input, and z is the new variable.

- **Data Re-scaling**:
  Data comes in different formats, shapes, and sizes. Before proceeding to building the model, data must be unified in terms of scales. Hence, this step is used to ensure all data samples falls within the same scale.

Model Learning

One of the ML techniques is used to build a model that derives patterns in data, describes relationships between data samples, and is able to make predictions for future unseen inputs.

Model Testing and Evaluation

After the model is constructed, it needs to be tested and evaluated in order to determine its accuracy and hence, usability for future reference. The testing stage is done using new data - that the model has not seen before. Whereas the evaluation stage, it is obtained using several metrics corresponding to the type of the built model.

1.4.3 Clustering

Clustering is an unsupervised ML technique whose aim is to segregate groups with similar traits and assign them into clusters by dividing data points into a number of groups (clusters) such that data points in the same groups are more similar to other data points in the same group than those in other groups. Several clustering techniques have been developed, from which we distinguish the three following ones:

Exclusive Approach (K-means)

It is a centroid-based algorithm that is the most commonly used. It opts to minimize the variance of data points within the same cluster.

Overlapping Approach (Fuzzy C-means)

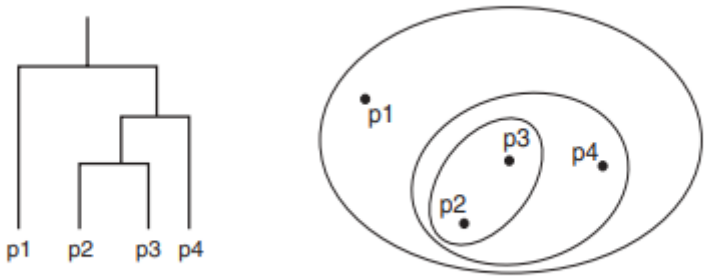
It is a soft method that allows data to belong to more than one cluster. Each cluster has a set of membership coefficients that are based on the degree of cluster membership.

Agglomerative Approach (Hierarchical Clustering)

It is a clustering method used to group data sample based upon how similar they rae to each other. Since it build clusters upon similarity, our work is based on it.

1.4.4 Hierarchical Clustering

Hierarchical clustering is an algorithm that builds hierarchy of clusters. It starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. This process is repeated until there is only a single cluster left.



(a) Dendrogram. (b) Nested cluster diagram.

Figure 1.2: A hierarchical clustering of four points shown as a dendrogram and nested clusters [27].

1.4.5 Dendrogram

Hierarchical clustering results can be shown by a dendrogram. Observations that fuse at the bottom are similar, whereas those at the top are quite different. A conclusion is drawn based on the location on the vertical rather than the horizontal axis.
The construction of a dendrogram consists of the following steps:

1. Begin with n observations and a measure of all the $\frac{n(n-1)}{2}$ pairwise distances. Taking each observation as a cluster.

2. For $i = n, n-1, ..., 2$:

   - Examine all the pairwise inter-cluster distances among the i clusters and identify the pair of clusters that are most similar (smaller distance — closer).
   - Fuse these two clusters. The dissimilarity between these two clusters indicates height in dendrogram where fusion should be placed.
   - Assign each observation to the cluster whose centroid is closest using a distance metric.

The following figure illustrates a dendrogram obtained from six data samples $(1, ..., 6)$. The closest samples are then grouped under one cluster; cluster number 7 fusing clusters 2 and 3. This procedure is repeated as explained above until obtaining one cluster (cluster 11).

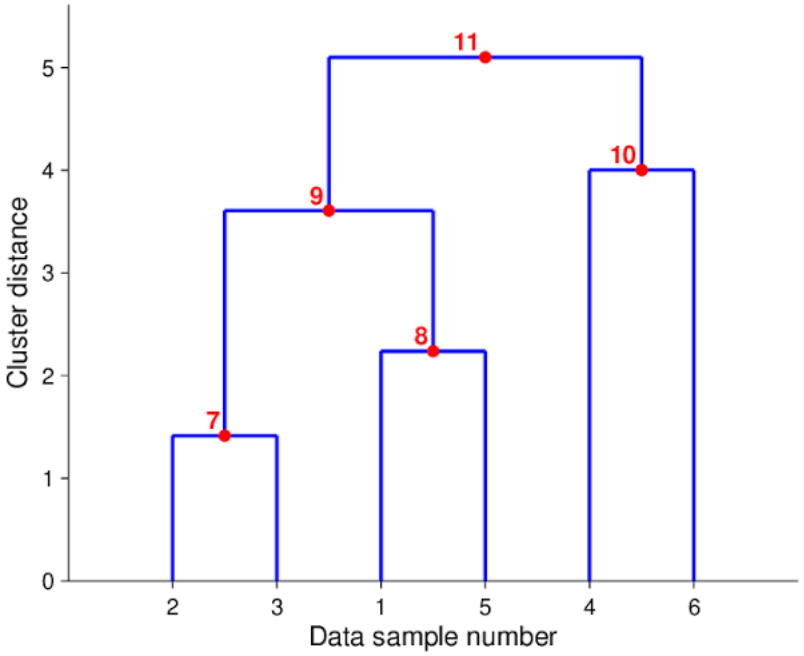


Figure 1.3: An example dendrogram for hierarchical clustering [28].

A threshold can be established to determine the number of classes wanted. For example, if 5 classes are needed, the cluster distance must be chosen around 2, forming classes $7, 1, 5, 4$, and 6. If the cluster distance is chosen to be less than 1, six lasses are obtained. A key aspect of hierarchical clustering process is how to compute the distance between two existing clusters in order to make a decision on how to group the closest ones together. In the following sections, distance and similarity measurements methods are discussed.

1.5 Distance Measurements Methods

Distance measures are the fundamental principle for classification. In data science, the similarity measure is a way of measuring how data samples are related or close to each other. This is often used in clustering when similar data samples are clustered in one group. The similarity measure is expressed as a numerical value describing how similar the two points (x, y) are. This value gets larger when the two points are more alike. This metric defines how the similarity of two points (x, y) is computed, and will affect the shape of the resulted clusters.

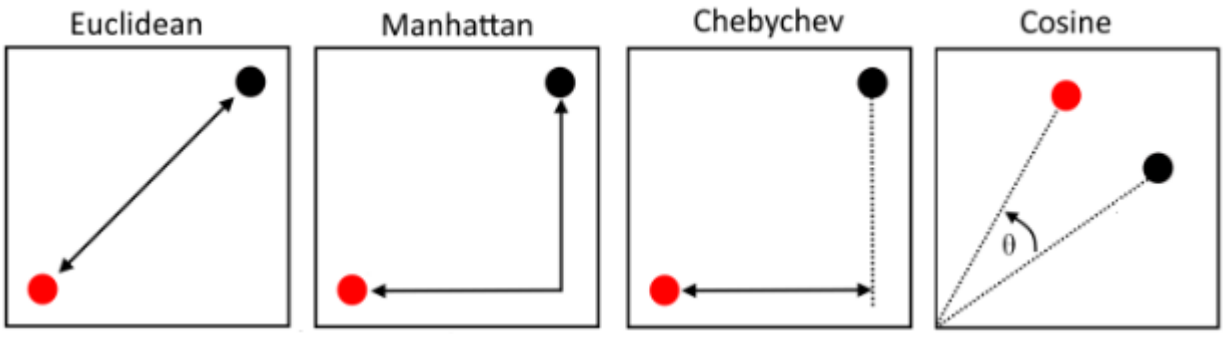


Figure 1.4: The most common distance measures [29].

The most distance measures, which are illustrated in Figure 1.4, are given by:

- $L2$ norm, Euclidean distance: which is the most common distance function. This represents the smallest distance between each pair of points in the sets P and Q. It is defined as:

$$d_{euc}(P,Q) = \|P - Q\| = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{1.1}$$

- $L1$ norm, City Block, Manhattan, or taxicab distance: it is very useful in measuring the distance between two streets in a given city, where the distance can be measured in terms of the number of blocks that separate two different places. This method was created to solve computing the distance between source and destination in a given city where it is nearly impossible to move in a straight line. It is expressed as :

$$d_{man}(P,Q) = \|P - Q\|_1 = \sum_{i}^{n} \|p_i - q_i\| \tag{1.2}$$

- $L\infty$ norm, Chebychev distance: it can be determined as the sum of absolute differences of their 2-dimensional coordinates. It is widely used in electronic computer-aided manufacturing applications such as drilling machines.

$$d_{cheb}(P,Q) = max|p_i - q_i| = |p_1 - q_1|or|p_2 - q_2|or....or|p_n - q_n| \tag{1.3}$$

- Cosine distance: this is widely used in information retrieval systems. It is measured as follows:

$$d_{cos}(P,Q) = 1 - cos(P,Q) = 1 - \frac{P.Q}{\|P\|.\|Q\|} = 1 - \frac{\sum_{i=1}^{n} P_i.Q_i}{\sqrt{\sum_{i=1}^{n} P_i^2}.\sqrt{\sum_{i=1}^{n} Q_i^2}} \tag{1.4}$$

In order to determine the closest cluster, different algorithms can be implemented depending on the link or augmentation criterion. The most common linkage criteria are the single link, the average link, the complete link, and Ward's link [30].

- In the single link algorithm, the cluster distance is determined by the closest observations (pair of two points) belonging to two different clusters.
- In the complete link algorithm, the distance between two clusters is determined by the two most distant points between two different clusters (furthest apart).
- For the average link, the distance between two clustered is averaged over all the distances between all possible point combinations between two clusters.
- Ward's link utilizes the distance between a central point in each cluster.

A common default is to use Ward's method which tends to result in balanced clusters. The complete linkage methods also yields to similar clusters. However, single linkage and complete linkage tends to result in many singletons and few large clusters.

1.6 Similarity Measures using Jensen-Shannon Divergence

In statistics and related fields, a similarity measure is a real-valued function that quantifies the similarity between two objects. Two of the used similarity metrics used in information retrieval are the Kullback-Leibler and Jensen-Shannon Divergence. Before getting into the latter methods, the entropy is first defined.

Entropy of a discrete random variable X, with distribution P, is a measurement of the amount of information required on the average to describe that variable. It is the most important metric in information theory as it measures the uncertainty of a given variable. Shannon defined the entropy H of a discrete random variable X with probability mass function $P(x)$ as [31]:

$$H(X) = -\sum_{x \in X} P(x) log_b P(x) \tag{1.5}$$

In the previous equation, if we set $b = 2$ in the log expression, we can estimate the minimum value in bits necessary to encode all the information contained in X.
The relative entropy measures how distant two distributions are from each other. It is also referred to as the Kullback-Leibler divergence between two samples.

1.6.1 Kullback-Leibler Divergence

The Kullback-Leibler divergence (KLD) is a measure of how a probability distribution differs from another probability distribution. Classically, in Bayesian theory, there is some true distribution $P(X)$; to which the estimate with an approximate distribution $Q(X)$ is needed [32]. In this context, the KLD measures the distance from the approximate distribution Q to the true distribution P.
Mathematically, considering two probability distributions $P(x), Q(x)$ on some space X, the KLD from Q to P (written as $D_{KL}(P\|Q)$) is:

$$D_{KL}(P\|Q) = -\sum_{x \in X} P(x) \times Log(\frac{P(x)}{Q(x)}) \tag{1.6}$$

Letting $P(x)$ and $Q(x)$, $x \in X$, be two probability mass functions (i.e. discrete distributions). Then $D_{KL}(P\|Q) \geq 0$ with equality if and only if $P(x) = Q(x)$ for all x. From the above formula, we note that The KLD is not symmetric [33]; that is:

$$D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$$

As a result, it is also not a distance metric. The KLD can take on values in $[0, \infty]$. Particularly, if P and Q are the exact same distribution, i.e. $P = Q$, then $D_{KL}(P\|Q) = 0$.
An alternate approach is the Jensen-Shannon divergence, another method of measuring the similarity between two probability distributions. It is a symmetric and smoothed version of the KL divergence and can be used as a distance metric.

1.6.2 Jensen-Shannon Divergence

The Jensen-Shannon divergence (JSD) is a principled divergence measure which is always finite for finite random variables. It is a symmetrized and smoothed version of

the Kullback-Leibler Divergence [34].
Given a sample x, to measure how likely x is to occur in the ground truth distribution P as opposed to the simulation distribution Q. The likelihood-ratio (LR) will measure this by:

$$LR = \frac{P(x)}{Q(x)} \tag{1.7}$$

$LR > 1$ indicates that $P(x)$ is more likely while $LR < 1$ indicates $Q(x)$ is more likely.
To imrove calculations over dataset, we take the sum of logs, as follows:

$$log(LR) = \sum_{i=1}^{m} log(\frac{P(x_i)}{Q(x_i)}) \tag{1.8}$$

JSD offers a smoother and symmetric approach by:

$$D_{JS}(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M) \tag{1.9}$$

where $M = \frac{1}{2}(P(x) + Q(x))$.
Using the Shannon entropy H from equation 1.5, the following equation is obtained:

$$D_{JS}(P\|Q) = H(\frac{P+Q}{2}) - \frac{H(P) + H(Q)}{2} \tag{1.10}$$

Equation 1.10 shows that the JSD is equivalent to the entropy of the mixture minus the mixture of the entropy.

The JSD approach has been used in several image processing applications. It has been used in edge detection [35], to select regions of interest in backgrounds [36] [37], and in image segmentation [38] [39]. Compared with the traditional information divergence, the Jensen-Shannon divergence is more accurate in measuring the similarity. [40]

1.7 Conclusion

Chapter 1 provided an overview of machine learning methods on medical image diagnosis. This chapter first gave some terminology about pulmonary diseases and their diagnosis. Statistics about lung diseases were presented in section 1.2, followed by detection and diagnosis approaches in section 1.3. Within this section, X-ray images were introduced in subsection 1.3.1, and in subsection 1.3.2, state-of-the-art algorithms were presented. The following section (1.4) covered the contribution of machine learning algorithms in assisting diagnosis using x-ray imaging. In 1.4.4, hierarchical clustering was the main focus. Then, dendrograms were detailed as a visualization of hierarchical clustering in 1.4.5. Distance and similarity measurements methods were presented in section 1.5, explaining different distance measurements and linkage methods used to construct our clusters. Finally, in subsections 1.6.1 and 1.6.2, Kullback-Leibler and Jensen-Shannon Divergence methods were tackled respectively.

Chapter 2

Similarity and Distance Measures for Hierarchical Clustering

2.1 Introduction

Many CAD systems rely on artificial neural networks (ANN) to classify images. However, not all tasks require supervised learning or neural networks. With an unsupervised approach, groups or clusters of images are determined without being constrained to a fixed feature. Determining data-driven groups without using a priori knowledge about labels or categories is the end goal of unsupervised clustering. The challenge of using different unsupervised clustering methods is that it will result in different partitioning of the samples. Thus, different groupings are formed since each method implicitly impose a structure on the data. The distance between samples (or the intrinsic relationships) can be measured with a distance metric (such as Euclidean distance), and stored in a so-called dissimilarity matrix. The distance between groups of samples can then be computed using the linkage type (for hierarchical clustering).
The proposed method consists of three steps: exploratory data, pre-processing, and classification step. Each method will be detailed in this chapter.

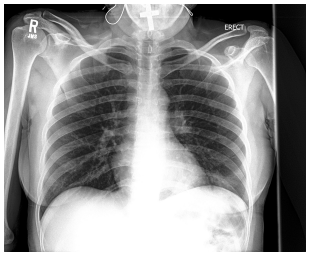
2.2 Exploratory Data

Lung disease detection is a classification task that requires chest X-ray images as input. The dataset used is the CheXphoto dataset [41]. It is publicly displayed and is created by a team of researchers from Stanford University, USA. It contains a total of 952 image of chest X-rays. The dataset has 14 labels; representing healthy cases, observations, and lung pathologies. The following table represents the acronyms used for each observation, as well as the number of images in which the observation is found.

Acronym	Observation	Number of Images
H	Healthy - No findings	138
EC	Enlarged Cardiomediastinum	327
Ca	Cardiomegaly	275
LO	Lung Opacity	378
LL	Lung Lesion	3
E	Edema	168
Co	Consolidation	114
Pn	Pneumonia	24
A	Atelectasis	306
Pt	Pneumothorax	24
PE	Pleural Effusion	242
PO	Pleural Other	3
F	Fracture	0
SD	Support Devices	321

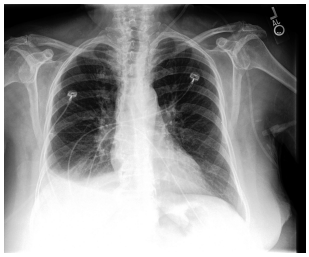
Table 2.1: CheXphoto Observations' acronyms.

Figure 2.1 shows a sample of images from CheXphoto dataset.

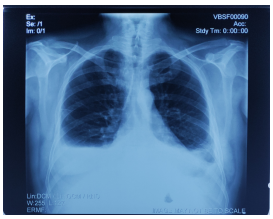


(a) Healthy

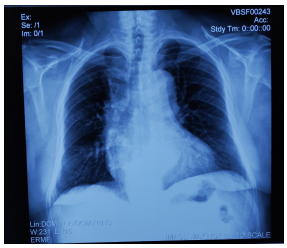
(b) Lung Opacity



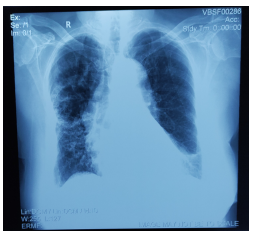
(c) Pleural Effusion



(d) Atelectasis



(e) Cardiomegaly



(f) Consolidation

Figure 2.1: Samples of XRays of different Lung Diseases.

The following table represents the first five lines of the tabulated data used in this study, where labels "F/L" and "AP/PA" have already been explained in table 1.1.

Path	Sex	Age	F/L	AP/PA	H	EC	Ca	LO	LL	E	Co	Pn	A	Pt	PE	PO	F	SD
CheX...	M	73	F	AP	0	1	1	1	0	0	0	0	1	0	0	0	0	0
CheX...	M	70	F	PA	0	0	0	0	0	0	0	0	0	0	0	0	0	1
CheX...	M	70	L		0	0	0	0	0	0	0	0	0	0	0	0	0	1
CheX...	M	85	F	AP	0	1	0	1	0	1	0	0	0	0	0	0	0	0
CheX...	F	42	F	AP	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2.2: First five rows of CheXphoto dataset.

2.3 Python Environment

Python [42] is an open source interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant white-space. It provides constructs that enable clear programming on both small and large scales. Python features a dynamic type system and automatic memory management. It

supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library. While implementing this project, several libraries were used.

Pandas

Pandas is a software library written for Python programming language used for data manipulation and analysis. It offers data structures and operations for manipulating numerical tables and time series. It is one of the most used tool in Machine Learning for cleaning and manipulating data.

NumPy

It is a library for Python that supports large, multi-dimensional arrays and matrices, along with high level mathematical functions. It is most used in domains of linear algebra, Fourier transform, and matrices manipulation.

Matplotlib

It is a comprehensive library for creating visualizations in Python.

PIL

Python Imaging Library, or PIL, is used to adding image processing capabilities to the Python interpreter. It adds support for opening, manipulating, and saving different image file formats.

SciPy

SciPy, standing for Scientific Python, is a scientific computation library. It provides utility functions for optimization, statistics and signal processing. It is designed to give more extensions of finding scientific mathematical formulae like Euclidean distance, linkage, Jensen-Shannon Divergence, and dendrograms.

2.4 Data Pre-processing

Before the construction of the classifier, data must be preprocessed to ensure that images are comparable in color, size, and value range. This study is conducted without taking into account both gender and age. Hence, the second and third columns of table 2.2 are ignored. In addition, the primary focus is the study of frontal images, and those that are PA since they provide a clearer view of lung tissues. This criteria has allowed us to eliminate lateral and AP images from the used dataset.

By analyzing the labels, it is necessary to note that there are some that represent symptoms - not pathologies. For example, edema, which refers to a swelling caused by fluid trapped in lungs' tissues, is considered as a result of an underlying disease. Taking this into consideration, the dataset observations are reduced to seven classes; healthy,

pneumonia, pneumothorax, lung opacity, atelectasis, pleural effusion, and enlarged cardiomediastinum.

Scanning through the images to be used, it was found that many of them represent a positive case for many pathologies. For instance, the first row in table 2.2 represents a chest x-ray of a patient with enlarged cardiomediastinum, atelectasis, and pleural effusion. In order to avoid this overlapping, only images that represent a single case were considered. The new dataset can be summarized in the following table:

Classes	Number of images
Healthy - no findings	8
Pneumonia	7
Pneumothorax	4
Atelectasis	5
Enlarged cardiomediastinum	7
Lung Opacity	3
Pleural Effusion	4

Table 2.3: Tabulated dataset resulted from preprocessing CheXphoto original dataset.

In the second step of pre-processing data, all images are converted from RGB (Red, Green, and Blue) to Grey scale images in order to reduce computations as the original images are in black and white. This results in data stored as matrices (2-D data).
Then, pixel values are normalized between $[0, 255]$ in order to avoid biased calculations. Finally, only images of the same size are used to ease computations of distances and similarities which require data to be of the same size.

2.5 Classifier Construction

A cluster is an area of density in the feature space where samples from the domain (observations) are closer to the cluster than other clusters. In hierarchical clustering, clusters are formed from a tree-type structure based on the hierarchy. New clusters are formed using the previously formed ones. In this work, agglomerative (bottom-up) approach is followed.

Many algorithms use similarity or distance measures between samples in the feature space in an effort to discover dense regions of observations. Our approach suggests testing various similarity and distance measurements to form clusters that are represented by dendrograms.

This work's approach consists of measuring the distance or similarity between images of each class. Then, constructing clusters based on the distances or similarities. These clusters are visualized by a dendrogram. Three metrics have been proposed to measure the similarity/distance between images, whereas the 'Ward' linkage has been used when constructing the dendrograms.

2.5.1 Euclidean Distance

In this study, in order to construct clusters of images, Euclidean distance approach was first considered. The Euclidean distance between two points is the length of a line segment between the two points. It is the prototypical example of the distance in a metric space, and obeys all the defining properties of a metric space.
Properties of a metric are, for every three points p, q, and r: [43]

- symmetric, where the distance between two points does not depend on which of the two points is the start and which is the destination; i.e. $d(p,q) = d(q,p)$,
- positive, meaning that the distance between every two distinct points is a positive number, while the distance from any point to itself is zero; i.e. $d(p,q) > 0$ and $d(p,p) = 0$,
- and obeys the triangle inequality: , $d(p,q) + d(q,r) \geq d(p,r)$.

As a reminder, the Euclidean distance is calculated by equation 1.1,

$$d_{euc}(P,Q) = \|P - Q\| = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Letting x, y be two $M \times N$ images, $x = (x^1, x^2, ..., x^{MN})$, $y = (y^1, y^2, ..., y^{MN})$, where x^{kN+l}, y^{kN+l} are the gray-scale levels at location (k,l). I our approach, row-wise distances and column-wise distances are considered separately. Hence, the row-wise and column-wise Euclidean distances are given by:

$$d_{Euc} = \sum_{k=1}^{M}(x^k - y^k)^2 \tag{2.1}$$

$$d_{Euc} = \sum_{k=1}^{N}(x^k - y^k)^2 \tag{2.2}$$

The algorithm used in this method is represented in the flowchart below. The aim is to calculate the distance between the healthy case and each of the six pathologies, both row-wise and column-wise, to obtain a dendrogram clustering the distances between each class.

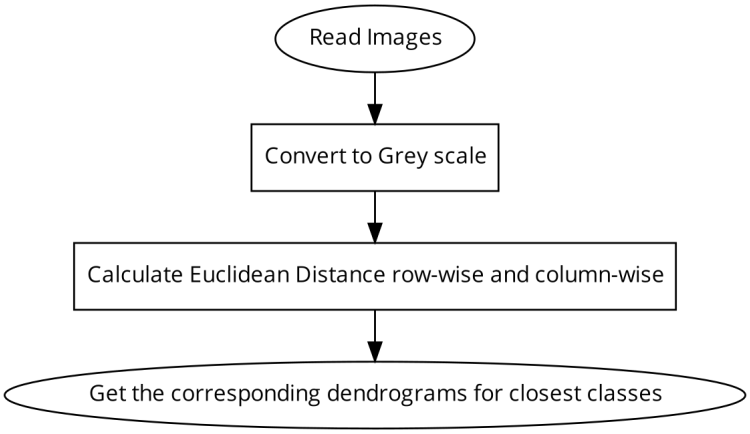


Figure 2.2: Euclidean distance method flowchart.

In Python, this can be obtained using the function `pdist` that iterates over rows and columns to calculate the Euclidean distance, as well as the `linkage` function which results in the 'Ward' distance needed to draw the dendrogram.

2.5.2 Cosine Distance

In data analysis, cosine similarity is a measure of similarity between two sequences of numbers. As a definition, the sequences are viewed as vectors in an inner product space, and the cosine similarity is defined as the cosine of the angle between them, i.e. the dot product of the vectors divided by the product of their lengths. It follows that the cosine similarity does not depend on the magnitudes of the vectors, but only on their angle. The expression has already been given by equation 1.4

$$d_{cos}(P,Q) = 1 - \frac{\sum_{i=1}^{n} P_i.Q_i}{\sqrt{\sum_{i=1}^{n} P_i^2}.\sqrt{\sum_{i=1}^{n} Q_i^2}}$$

The cosine distance is bounded by $[-1, 1]$. For example, two proportional vectors have a cosine similarity of 1, two orthogonal vectors have a similarity of 0, and two opposite vectors have a similarity of -1. The cosine similarity is particularly used in positive space, where the outcome is bounded in $[0, 1]$.

The most noteworthy property of cosine similarity is that it reflects a relative, rather than absolute, comparison of the individual vector dimensions. The measure is thus most appropriate for data where frequency is more important than absolute values.

Prior to using the Euclidean distance, we modify our flowchart to obtain the cosine distance between each pairs. The rest of the procedure remains unchanged. Whereas in Python, it is only needed to add the attribute `'cosine'` in the `pdist` function to obtain the cosine distance.

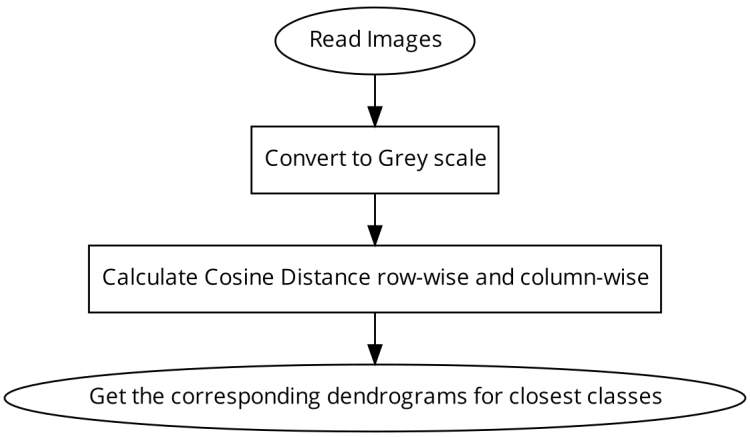


Figure 2.3: Cosine distance method flowchart.

2.5.3 Jensen-Shannon Divergence

In probability theory and statistics, the Jensen–Shannon divergence (JSD) is a method of measuring the similarity between two probability distributions. It is based on the Kullback–Leibler divergence (KLD), with some notable differences, including that it is

symmetric and it always has a finite value. On the countrary to KL divergence, it determines in a very direct way a metric; it is the square of a metric [44]. Hence, the square root of the JSD is a metric often referred to as Jensen–Shannon distance.
The JS divergence expression has already been mentioned in equation 1.9, that is:

$$D_{JS}(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M)$$

where $M = \frac{1}{2}(P + Q)$.
The JSD is bounded by 1 for two probability distributions given that base 2 is used, i.e.

$$0 \leq D_{JS}(P\|Q) \leq 1$$

. JSD is implemented in Python using the function `jensenshannon` with attribute `'base=2'`. The following flowchart represents the algorithm proposed for this approach.

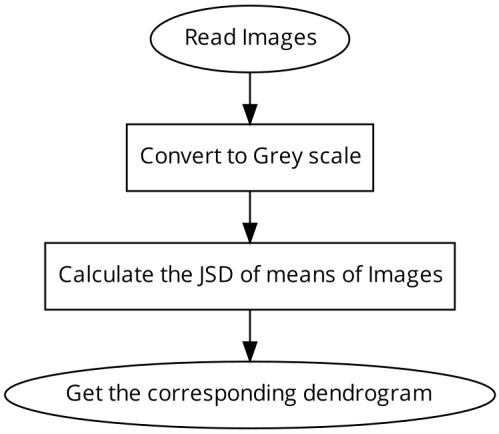
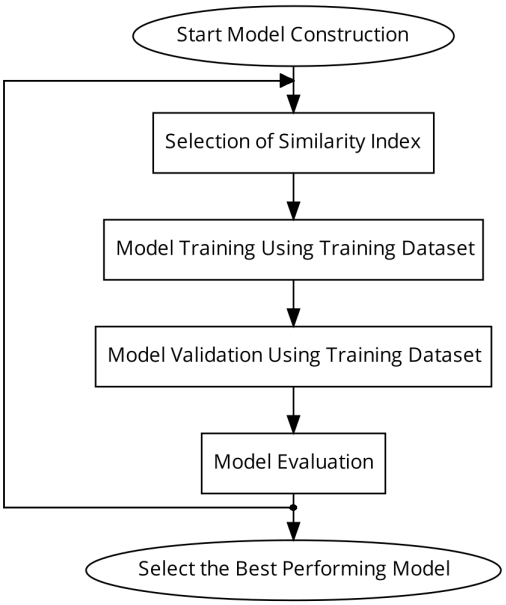


Figure 2.4: Jensen-Shannon Divergence method flowchart.

2.6 Model Validation and Evaluation

When using machine learning models, it is insufficient to rely on the training set results only - validation must be performed. In the process of model validation, the model is subjected to data that has not been presented before in the training phase. Tho goal is the ability of the model to curry out correct and reliable clustering even when new data is entered. The following flowchart illustrates the steps of constructing a model.



Evaluation of identified clusters is subjective and may require a domain expert, although many clustering-specific quantitative measures do exist. Typically, clustering algorithms are compared academically on synthetic datasets with pre-defined clusters, which an algorithm is expected to discover.
The ultimate goal of a clustering algorithm is to achieve high intra-cluster similarity and low inter-cluster similarity. In other words, we want data points in the same cluster to be as close to each other as possible. Whereas the distance between different clusters needs to be as high as possible.
In order to visualize how well the model is doing on the test set - a set that the proposed architecture has not seen during training - we use metrics. There are plenty of metrics used to evaluate ML models, for our purpose, we chose the three widely used techniques that are: average accuracy, purity, and F-1 score.

2.6.1 Confusion Matrix

A confusion matrix is a square matrix that is used as a way of visualizing the performance of our prediction model. Each entry in a confusion matrix denotes the number of predictions that were made by the model on the test set where it classified the classes correctly or incorrectly. The number of rows in a confusion matrix is defined by the number of classes.

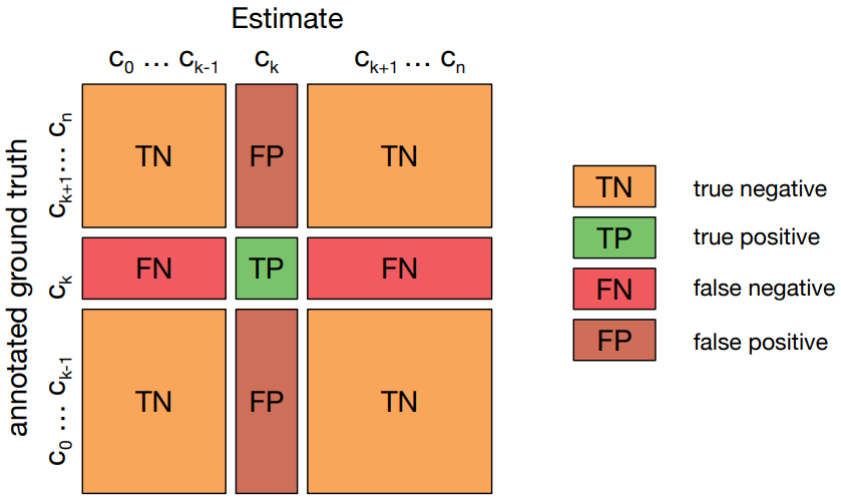


Figure 2.5: Confusion Matrix for n classes [45].

From the confusion matrix in figure 2.5, when considering the class k ($0 \leq k \leq n$), we can derive four important values for each class i that are used for computing the metrics mentioned before. If we consider class i to be positive (1) and the other $n-1$ classes to be negative (0), then:

- True Positive Rate (TPR), also known as the sensitivity, refers to the number of predictions where the classifier predicts the class i correctly positive.

$$Sensitivity = TPR = \frac{TP}{TP + FN} \tag{2.3}$$

- True Negative Rate (TNR), also referred to as the specificity, refers to the number of predictions where the classifier correctly predicts the negative class i as negative.

$$Specificity = TNR = \frac{TN}{TN + FP} \tag{2.4}$$

- False Positive Rate (FPR) refers to the number of predictions where the classifier incorrectly predicts the other $N-1$ class as class i.

$$FPR = \frac{FP}{TP + FN} = 1 - Specificity \tag{2.5}$$

- False Negative Rate (FNR) refers to the number of predictions where the classifier incorrectly predicts class i classes as one of the $N-1$.

$$FNR = \frac{FN}{TP + FN} \tag{2.6}$$

2.6.2 Average Accuracy

It gives the overall accuracy of the model, meaning the fraction of the total samples that were correctly classified by the classifier.

$$ACC_{avg} = \frac{1}{n} \sum_{i=1}^{n} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{2.7}$$

2.6.3 F1-score

The harmonic average of the precision and recall, it measures the effectiveness of identification when just as much importance is given to recall as to precision for each class.

$$F1 - score = \frac{precision \times recall}{precision + recall} \tag{2.8}$$

Precision

It tells us what fraction of predictions as a positive class were actually positive. To calculate precision we use:

$$precision = \frac{TP_i}{TP_i + FP_i} \tag{2.9}$$

Recall

It is the ability of the model to find correct predictions per class.

$$recall = \frac{TP_i}{TP_i + FN_i} \tag{2.10}$$

2.6.4 Purity

The purity represents the number of correctly matched class and cluster labels divided by the number of total data points. Each cluster is assigned with the most frequent class labels. It is defined by [46]:

$$purity = \sum_i \frac{N_i}{N} p_i \tag{2.11}$$

Where:

- N is the total number of objects,
- $N_i = \sum_{j=1}^{C} N_{ij}$ be the total number of objects in cluster i, and N_{ij} be the number of objects in cluster i that belong to class j,
- $p_i = max_j p_{ij}$ where $p_{ij} = \frac{N_{ij}}{N_i}$ is the empirical distribution over class labels for cluster i.

The purity is bounded between 0 (bad clustering) and 1 (good clustering). However, we can easily achieve a purity of 1 by putting each object into its own cluster; this measure does not penalize for the number of clusters. In general, purity increases as the number of clusters increases. For instance, if we have a model that groups each observation in a separate cluster, the purity becomes one.

2.7 Conclusion

This chapter covered the proposed approaches studied in this work. First, the tools used were introduced starting by the CheXphoto dataset in section 2.2, and the Python programming language in 2.3. Since the images were not comparable in color, dimension, and pixel value range, pre-processing steps have been applied. These modifications were

mentioned in section 2.4. In section 2.5, the three different methods used to build clusters have been discussed and each method's flowchart was presented. Finally, section 2.6 shed light on the criteria used to evaluate the models proposed.

Chapter 3

Application, Results, and Discussion

3.1 Introduction

This chapter sheds the light on the validation of the models built in the previous chapter. The three methods were tested on data collected from the CheXphoto dataset. First, clusters were built for each model. Then, the test set is used to predict corresponding clusters for each sample. Finally, these models were evaluated using previously-mentioned metrics.
In the following section, each method's resulting dendrograms are presented. Clusters are henceforth formed based upon a threshold that separates the healthy class in an individual cluster. Then, two evaluation metrics are used to test the accuracy of the clustering, which are the Average Accuracy and purity. Results and findings for each method are compared and discussed at the end of this chapter.

3.2 Results and Findings

In this section, the resulted dendrograms are discussed in order to form clusters. The main criterion to determine the number of clusters for each method is the threshold that isolates the healthy class (denoted H) from the faulty classes. Then, each method's cluster tested and evaluated using the average accuracy, F-1 score, and purity.

3.2.1 Euclidean Distance

The Euclidean distance method proposed consists of constructing two dendrograms; one for the row-wise distances, and the other for the column-wise distances. After building the model from the flowchart illustrated in Figure 2.2, Figures 3.1 and 3.2 represent the outputs obtained.

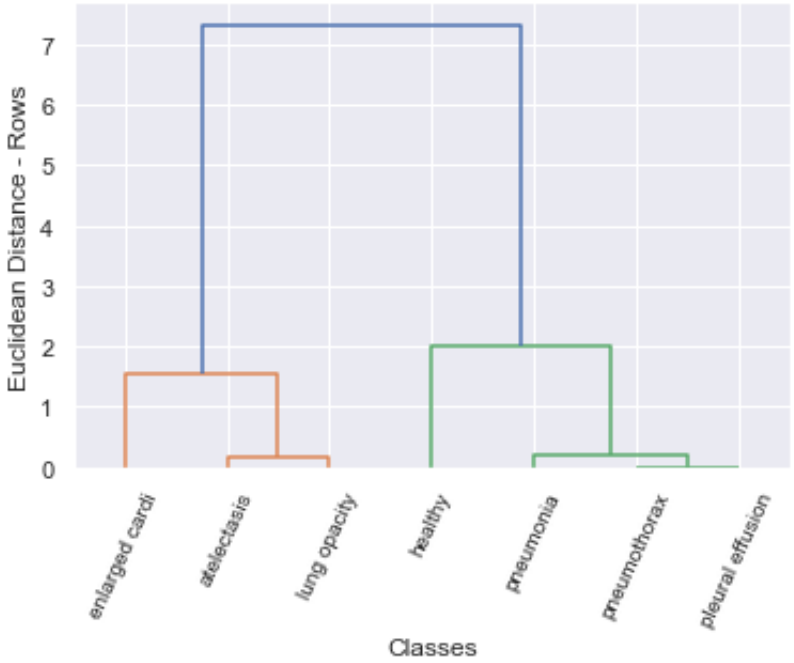


Figure 3.1: Dendrogram representing Row-wise Euclidean Distance of each class.

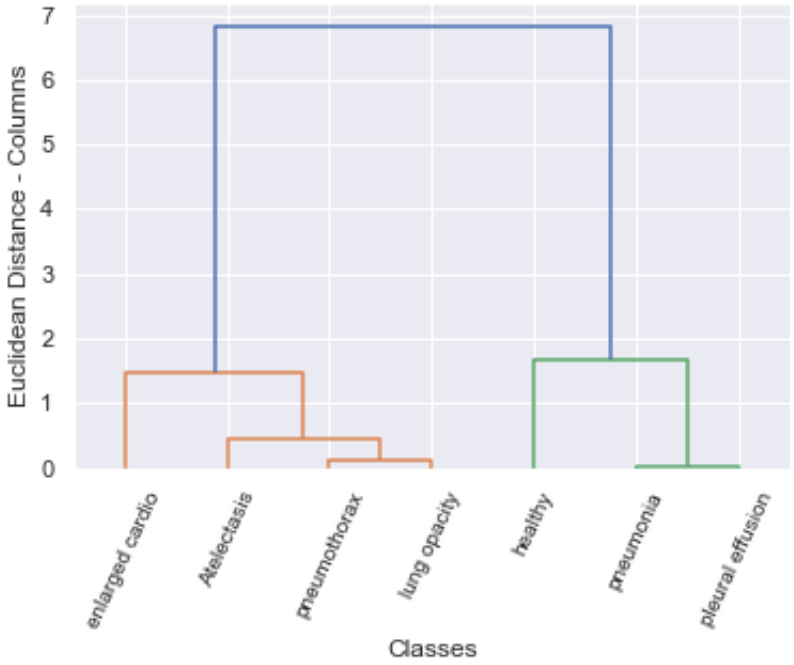


Figure 3.2: Dendrogram representing Column-wise Euclidean Distance of each class.

We note that the arrangement of the classes is not the same when considering row-wise and column-wise distances. From Figure 3.1, it can be seen that the classes Pleural Effusion and Pneumothorax (located at the extreme left) have a smaller Euclidean distance with the class Pneumonia. However, in Figure 3.2, the class Pneumothorax is clustered far away to the right on Pneumonia and Pleural Effusion. This difference in arrangement must be taken into consideration. We cannot consider only one dendrogram; there is a visible difference when considering column-wise or row-wise Euclidean distances.

In order to differentiate the healthy case, at least four clusters need to be formed. First, only four clusters are formed. However, two arrangements of clustering are possible; the first one is by taking the row-wise distance first, then using the column-wise distance to confirm the classification if the image has two or more possible classes. This clustering is explained in Table 3.1.

Cluster	Pathologies
1	H
2	EC
3	A/LO
4	P/Pt/PE

Table 3.1: Row-wise Euclidean Distance with four clusters.

In order to determine how accurate this classifier is, it needs to be tested and evaluated using previously discussed metrics. The testing procedure is provided by the flowchart below - for row-wise distance:

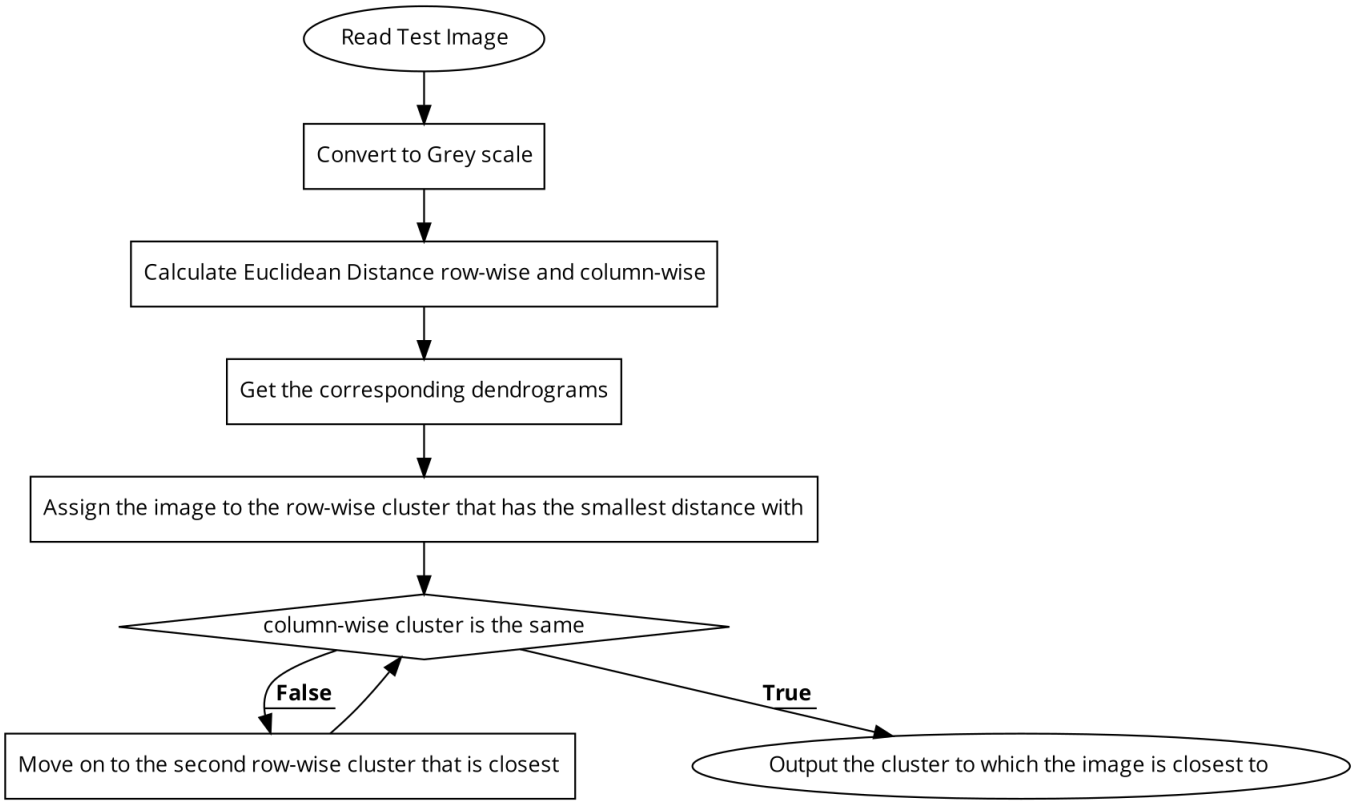


Figure 3.3: Testing procedure for the Row-wise Euclidean Distance method.

After obtaining the results of test images, opting to evaluate the model consists first of constructing the confusion matrix. For this clustering, it is given by:

Cluster	1	2	3	4
1	4	1	1	2
2	1	1	1	4
3	2	0	0	2
4	2	1	2	10

The average accuracy is obtained from Equation 2.7 as :

$$ACC_{avg} = 0.7218$$

Wheras the purity is obtained from Equation 2.11:

$$Purity = 0.5$$

And the F-1 score is calculated from Equation 2.8 :

$$F - 1score = 0.235$$

The second method takes column-wise distance first into consideration, then uses row-wise distance to confirm the cluster to which a test image belongs to. The clusters formed are:

Cluster	Pathologies
1	EC
2	A/Pt/LO
3	H
4	P/PE

Table 3.2: Columns-wise Euclidean Distance with four clusters considering.

The testing procedure for the column-wise distance is illustrated below.

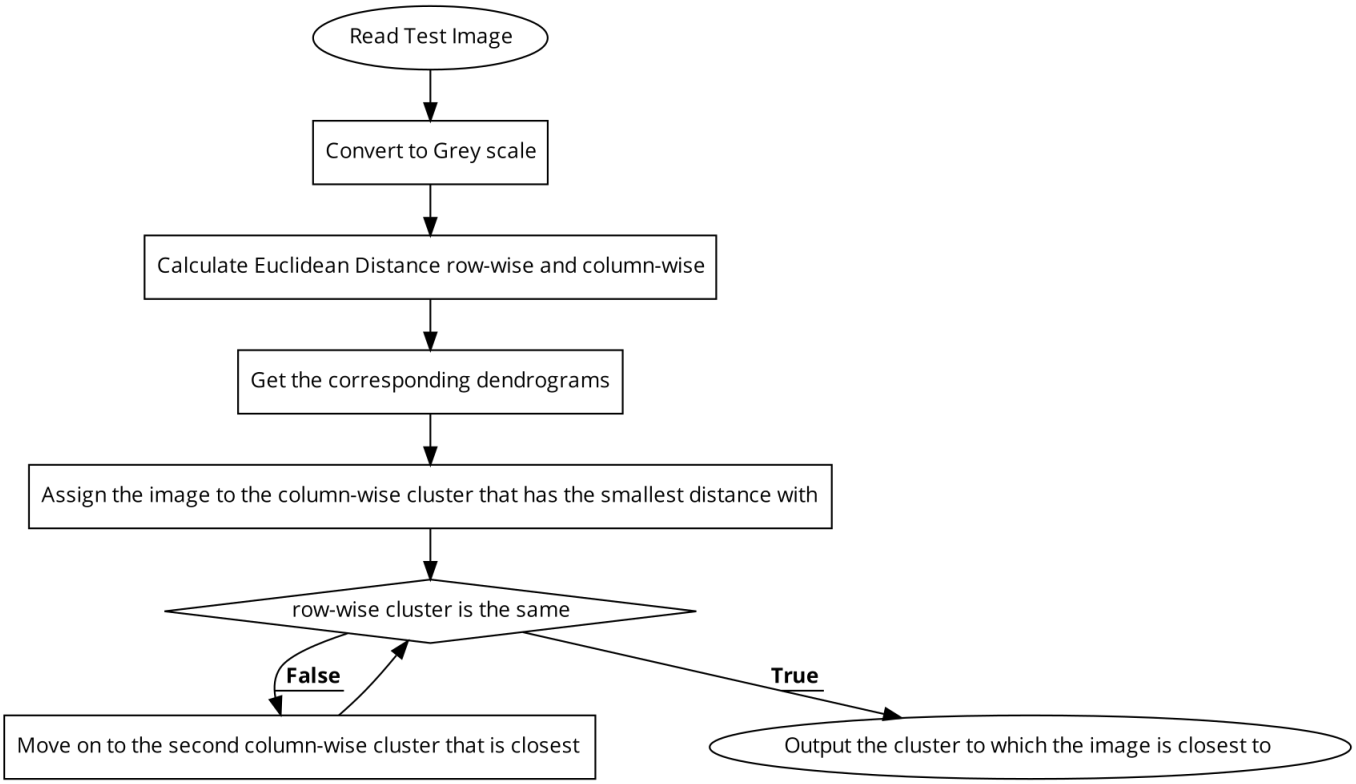


Figure 3.4: Testing procedure for the Euclidean Distance method with Column-wise distance priority.

This method allows the construction of the following confusion matrix:

Cluster	1	2	3	4
1	1	2	1	3
2	0	8	2	2
3	1	2	4	1
4	1	4	2	4

The average accuracy obtained is:

$$ACC_{avg} = 0.7237$$

Wheras the purity is:

$$Purity = 0.4474$$

and the F-1 score is:

$$F-1score = 0.0.203$$

By choosing to set five clusters, our classifier will have two possible clusters (one when taking rows distance into consideration first, and the other if the columns distance is considered first).Following the same procedure as above (for four clusters), the five possible clusters are tabulated below:

Cluster	Pathologies
1	EC
2	A/LO
3	H
4	P
5	Pt/PE

Table 3.3: Row-wise Euclidean Distance with five clusters.

Cluster	Pathologies
1	EC
2	A
3	Pt/LO
4	H
5	P/Pt

Table 3.4: Column-wise Euclidean Distance with five clusters.

First, let's consider Table 3.3. This results in the following confusion matrix:

Cluster	1	2	3	4	5
1	1	1	1	3	1
2	0	4	2	0	2
3	1	1	4	0	2
4	0	1	1	3	2
5	1	1	1	1	4

This matrix results in purity:

$$Purity = 0.4211$$

and average accuracy of:

$$ACC_{avg} = \frac{1}{5}(3.080) = 0.6160$$

and of the F-1 score of:

$$F-1score = 0.202$$

By considering the column-wise distance first, as shown in table 3.4, five clusters resulted differently. The corresponding confusion matrix is:

Cluster	1	2	3	4	5
1	1	1	1	1	3
2	0	1	2	2	0
3	0	0	5	0	2
4	1	0	2	4	1
5	1	0	4	2	4

This matrix results in average accuracy of:

$$ACC_{avg} = \frac{1}{5}(3.802) = 0.7604$$

and purity of:

$$Purity = 0.3947$$

Whereas the F-1 score is:

$$F - 1score = 0.181$$

The next step consists of setting the number of clusters to be six. The same procedure is repeated. The clusters are detailed in the tables below:

Cluster	Pathologies
1	EC
2	A
3	H
4	P
5	PE/Pt
6	LO

Table 3.5: Row-wise Euclidean Distance with six clusters.

Cluster	Pathologies
1	EC
2	A
3	Pt
4	LO
5	H
6	P/PE

Table 3.6: Column-wise Euclidean Distance with six clusters.

The confusion matrix obtained by table 3.5 is:

Cluster	1	2	3	4	5	6
1	1	1	1	3	1	0
2	0	1	2	0	0	2
3	1	0	4	0	2	1
4	0	0	1	3	2	1
5	1	0	1	1	4	1
6	0	0	0	0	2	1

The corresponding average accuracy, purity, and F-1 score for this matrix are:

$$ACC_{avg} = 0.7895$$

$$Purity = 0.3684$$

$$F - 1score = 0.169$$

Whereas for table 3.6, the matrix bellow is obtained.

Cluster	1	2	3	4	5	6
1	1	1	1	0	1	3
2	0	1	0	2	2	0
3	0	0	2	0	0	2
4	0	0	2	1	0	0
5	1	0	1	1	4	1
6	1	0	2	2	2	4

The corresponding average accuracy, purity, and F-1 score for this matrix are:

$$ACC_{avg} = 0.7864$$

$$Purity = 0.3421$$

$$F - 1score = 0.160$$

Finally, the number of clusters is set to the number of classes, i.e. seven. It is notable that the clusters are the same when considering row-wise distance or column-wise distance. Hence, only one confusion matrix is obtained, which is:

Cluster	1	2	3	4	5	6	7
1	4	0	1	1	1	0	1
2	1	3	2	0	1	0	0
3	0	1	2	1	0	0	0
4	1	0	0	1	1	0	1
5	0	0	2	0	1	0	0
6	2	0	0	0	2	1	0
7	1	3	1	0	0	1	1

The purity, F-1 score, and average accuracy are calculated. The results are:

$$ACC_{avg} = 0.8120$$

$$Purity = 0.3421$$

$$F - 1score = 0.236$$

3.2.2 Cosine Distance

By applying the algorithm shown by Figure 2.3, the following dendrograms are obtained:

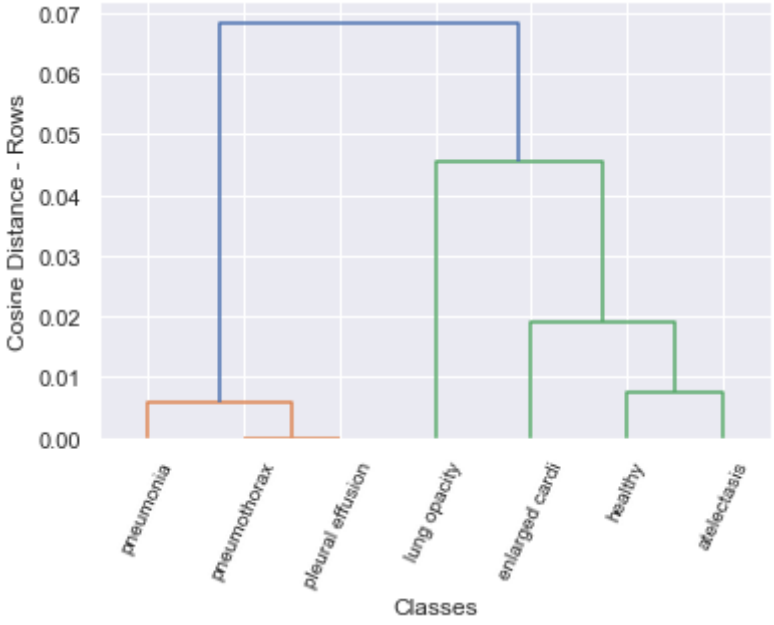


Figure 3.5: Dendrogram representing Cosine Distance with respect to rows of each class

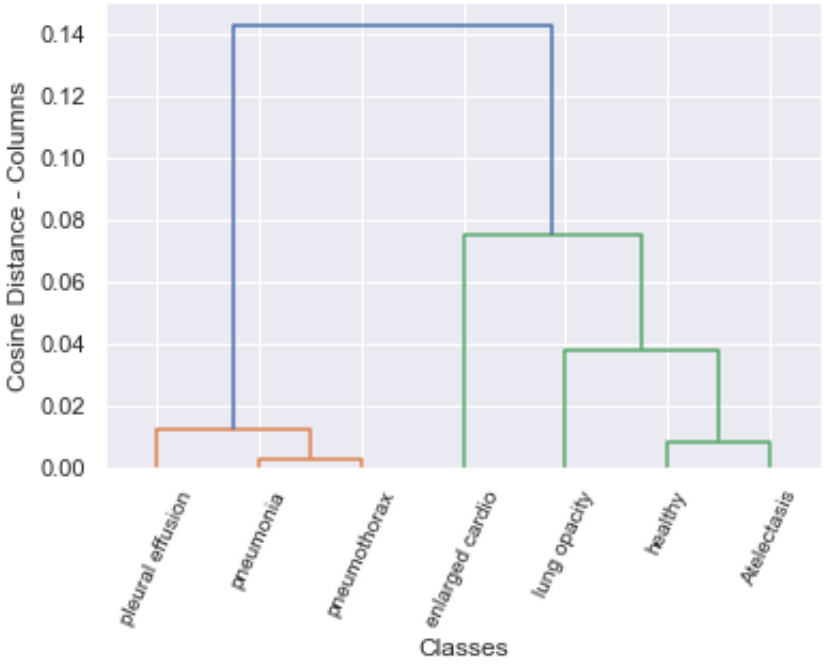


Figure 3.6: Dendrogram representing Cosine Distance with respect to columns of each class

Figures 3.5 and 3.6 show that when using cosine distance as a similarity index, both dendrograms have the same shape and arrangement of classes, but with different values. Hence, clusters can be formed without having to take row-wise distance and column-wise

distance independently - unlike when using the Euclidean Distance. Moreover, in order to be able to distinguish the healthy class, at least five clusters are formed:

Cluster	Pathologies
1	P/ PT/ PE
2	LO
3	EC
4	H
5	A

Table 3.7: Cosine Distance with five clusters.

Since the arrangement of classes is the same row-wise and column-wise, the following flowchart is considered in clustering test images:

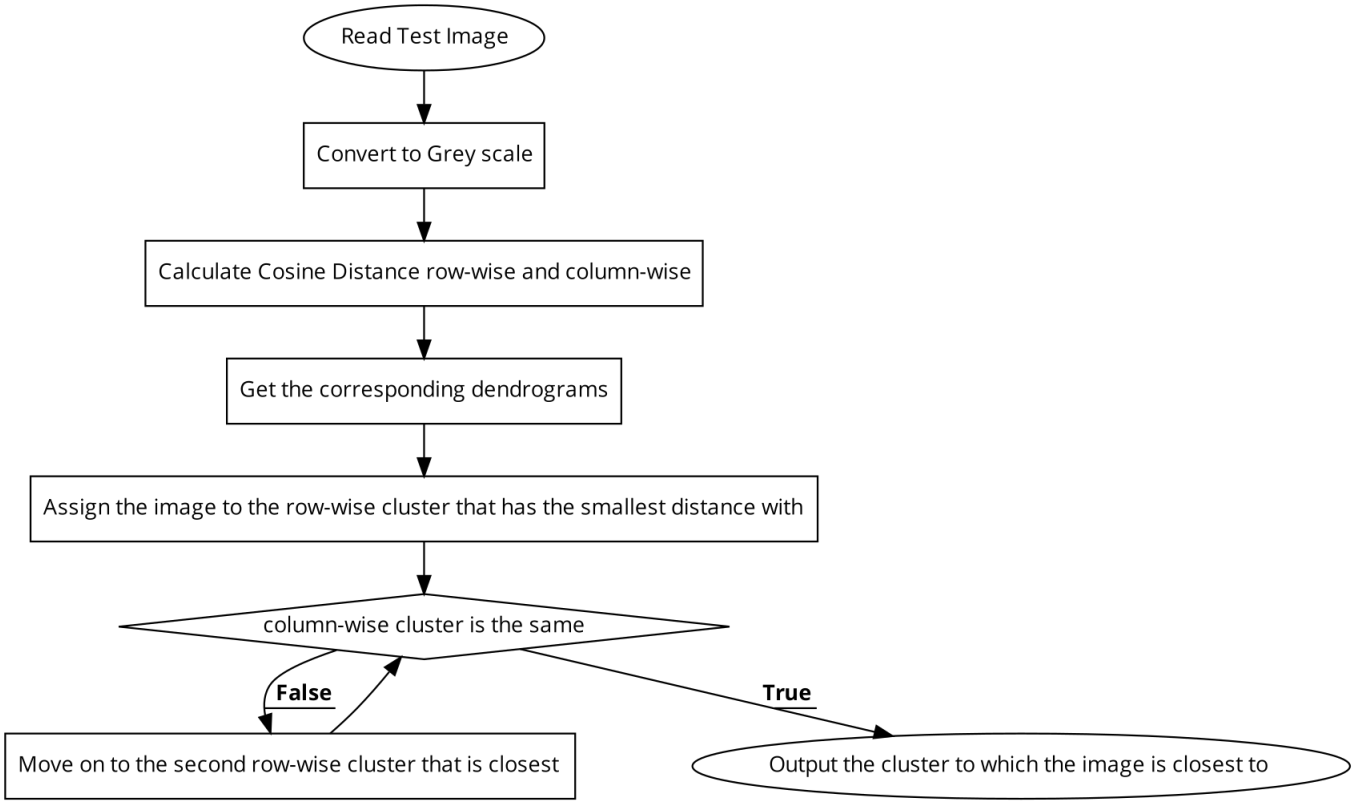


Figure 3.7: Testing procedure for the Cosine Distance method.

To test the proposed method, the confusion matrix and purity need to be calculated. First, the confusion matrix obtained from testing the model is:

Cluster	1	2	3	4	5
1	9	0	1	3	2
2	2	1	0	0	0
3	3	1	2	1	0
4	4	2	1	1	0
5	2	1	0	1	1

The average accuracy is obtained from the equation 2.7 as:

$$ACC_{avg} = 0.7313$$

and the purity of this cluster is calculated from equation 2.11 as:

$$Purity = 0.3684$$

and the F-1 score is obtained from Equation 2.8:

$$F - 1score = 156$$

The following table represents the case of choosing six clusters:

Cluster	Pathologies
1	P
2	Pt/PE
3	LO
4	EC
5	H
6	A

Table 3.8: Cosine Distance with six clusters.

From the table, the confusion matrix below is obtained:

Cluster	1	2	3	4	5	6
1	2	4	0	1	1	0
2	1	4	0	0	1	2
3	0	2	1	0	0	0
4	2	3	1	2	1	0
5	0	4	2	1	1	0
6	0	2	1	0	1	1

As proceeded before, the average accuracy, purity, and F-1 score are calculated as follow:

$$ACC_{avg} = 0.2895$$

$$Purity = 0.6341$$

$$F - 1score = 0.144$$

Finally, seven clusters are formed in by isolating each class by itself. The following confusion matrix is obtained:

Cluster	1	2	3	4	5	6	7
1	1	0	4	0	2	0	1
2	2	2	2	0	0	0	1
3	1	0	1	1	0	1	0
4	0	1	1	1	0	1	0
5	0	0	2	0	1	0	0
6	1	0	2	0	1	1	0
7	1	2	0	1	1	0	2

The average accuracy, purity, and F-1 score of this clustering are:

$$ACC_{avg} = 0.7791$$

$$Purity = 0.2368$$

$$F - 1score = 0.131$$

3.2.3 Jensen-Shannon Divergence

After testing the two previous distance indices, the JSD is tested at last. By applying the algorithm described by the flowchart in Figure 2.4, the dendrogram obtained is presented below:

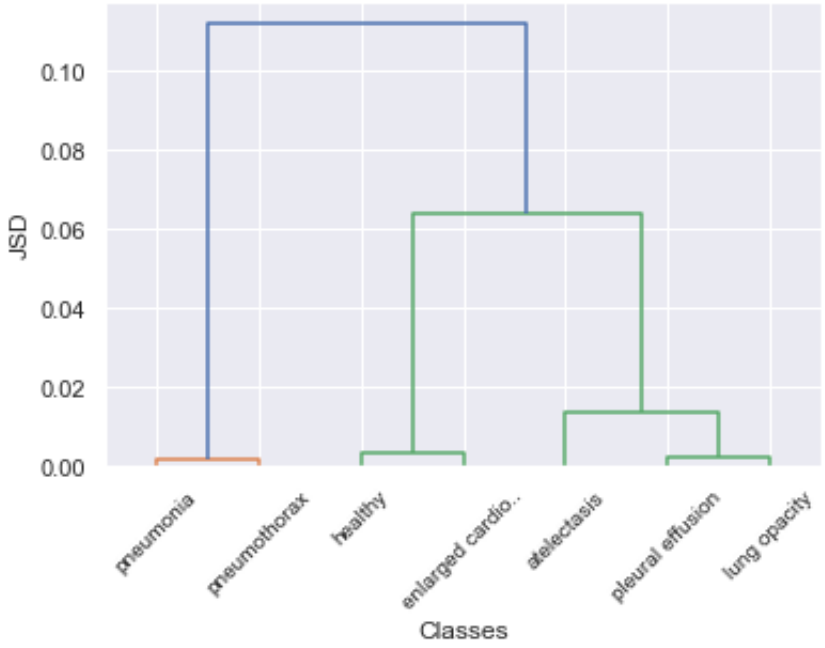


Figure 3.8: Dendrogram representing JSD of each class.

For the JSD similarity index, five clusters are formed in order to separate the healthy class from faulty observations. Clusters are illustrated in the table below:

Cluster	Pathologies
1	P/ PT
2	H
3	EC
4	A
5	PE / LO

Table 3.9: Jensen-Shannon Divergence with five clusters.

After setting the number of clusters to five, our model is tested. The average accuracy, F-1 score, and purity of this clustering are:

$$ACC_{avg} = 0.6736$$

$$Purity = 0.1842$$
$$F - 1score = 0.184$$

Setting the number of clusters to six (as shown in table 3.10 resulted in the following evaluation metrics' values:

Cluster	Pathologies
1	P/Pt
2	H
3	EC
4	A
5	PE
6	LO

Table 3.10: Jensen-Shannon Divergence six clusters.

$$ACC_{avg} = 0.6930$$
$$Purity = 0.0789$$
$$F - 1score = 0.079$$

Finally, when setting the number of clusters to the number of classes, the following results were obtained:

$$ACC_{avg} = 0.7360$$
$$Purity = 0.0789$$
$$F - 1score = 0.068$$

3.3 Discussion

The following table presents the comparison in the average accuracy and purity of each clustering method used. It summarized the results obtained in the previous section.

Number of clusters	Clustering method	Average Accuracy	Purity	F1-score
4	Euclidean distance - Rows	0.7218	**0.5**	**0.2657**
4	Euclidean distance - Columns	0.7237	0.4474	0.2031
5	Euclidean distance - Rows	0.6160	0.4211	0.2023
5	Euclidean distance - Columns	0.7604	0.3947	0.1812
5	Cosine distance	0.7313	0.3684	0.156
5	JSD	0.6736	0.1842	0.1840
6	Euclidean distance - Rows	0.7895	0.3684	0.1694
6	Euclidean distance - Columns	0.7864	0.3421	0.1601
6	Cosine distance	0.6341	0.2895	0.1444
6	JSD	0.6930	0.0789	0.0789
7	Euclidean distance	**0.8120**	0.3421	0.161
7	Cosine distance	0.7794	0.2368	0.1312
7	JSD	0.7360	0.0789	0.0682

Table 3.11: Hierarchical clustering technique with different similarity matrices.

The experimental results presented to evaluate the performance of the detection of the cardio-respiratory disease model have revealed several contrasting themes. The Euclidean distance with row-wise distances taken into consideration first has achieved the highest purity among all methods when four clusters have been set. Where seven clusters had achieved the highest average accuracy.

Column-wise Euclidean distance has often achieved a better accuracy than the row-wise Euclidean distance for the same number of clusters. However, the difference is very small.

When considering the cosine distance method, calculations are reduced due to the coherent resulted arrangements in the row-wise and column-wise - in contrary to Euclidean distance where each arrangement needs to be taken into consideration. From table 3.11, it can be seen that Cosine distance with five clusters has achieved the highest purity among other proposed cluster numbers, with a relatively good average accuracy.

In hierarchical cluster analysis, agglomeration of objects with almost the same levels of features values is better realized using Euclidean distance. However, if the purpose is to agglomerate objects with similar patterns that may vary by constant additive or multiplicative translation, then cosine distance achieves a better performance.

When using the square root of the JSD as a metric for image similarity, a very low purity was resulted in, regardless of the number of clusters chosen. Even though, JSD was able to cluster classes with around 70% accuracy, clusters' purity was very low (< 0.2) compared to the other methods. One of the JSD's advantages in this application is that it does not take into account the size of the images - unlike the Euclidean distance. This feature has allowed the use of a larger number of samples from the dataset that does not have the same size when testing this method.

There is always a trade-off when setting which method is best suitable. Euclidean distance offers a better purity, especially for a smaller number of clusters (four or five) compared to cosine distance. However, cosine distance offers less computations - is henceforth faster. Whereas JSD method takes images of different sizes - no need for re-scaling which results in a loss of information.

These methods have achieved acceptable scores in accuracy compared to the SOTA machine learning methods mentioned in Chapter 1 - the obtained accuracy range from 0.6160 to 0.8120. However, our approach offers faster computations and preservation of images' high dimensionality. Hence, it can be used in the primitive stages of detection with less than 30% inaccuracy.

3.4 conclusion

This chapter covered the results and findings of the proposed methods discussed in Chapter 2. First, finding were presented and clusters were extracted. Then, each method clustering was tested, and tabulated results in confusion matrices. Finally, evaluation metrics were computed in order to evaluate each method. These results were detained

in Section 3.2. In Section 3.3, a brief tabulated summary was set based upon evaluation metrics in order to compare each method's clusterings. The approach proposed in thesis was found meeting the previously mentioned gaps.

General Conclusion

This dissertation addresses the needs for methods and tools that assist physicians in diagnosing lung diseases. Our proposed approach is based upon using distance and similarity measures to construct clusters that separate lung conditions. The methods proposed neither need large computational powers nor large volume datasets. In fact, it preserves the high resolution size on original chest X-rays.

In Chapter 1, the motivation behind this study was explained by presenting the latest statistics regarding lung disease propagation and mortality both worldwide and in Algeria (section 1.2). Furthermore, the previous work done in lung disease detection has been discussed, shedding light on the limitations and challenges of the trendy machine learning models used nowadays. In the following sections, background theory of our proposed approaches were given. In section 1.5, distance measurements were first introduced, followed by similarity measure methods in section 1.6.

Chapter 2 focused on our contribution throughout this work. Section 2.2 and 2.3 introduced the main tools used to realize this project; the CheXphoto dataset and the Python programming language respectively. Needless to say, pre-processing steps were crucial before proceeding to build the classifier. These steps were discussed in section 2.4. In section 2.5, flowcharts of Euclidean distance, Cosine distance, and the Jensen-Shannon Divergence methods were explained. This section explained the core to our project. After building clusters, model evaluation was needed. Several metrics were opted to determine the quality of our proposed approach. These metrics were detailed in section 2.6.

The last Chapter, Chapter 3, covered tests and results of the proposed approaches after applying rigorous evaluation methods, and the results have been discussed and detailed in sections 3.2 and section 3.3.Findings demonstrated that these methods were able to cluster lung conditions and classify test images with relatively good accuracy using small datasets and preserved high dimensional images.

These approaches however, can be refined through additional studies in future works. Our approach consists of applying similarity and distance measurements to a publicly available dataset with high dimension images that is limited in size; future work is intended to collect more X-ray images, or use larger datasets to make it more representative. In addition, additional analysis are warranted of the effect of removing noisy clinical-assigned labels. Radiologists do not always agree on what diseases are present in a given X-ray. hence, different diseases pattern tends to look similar, further complicating the detection task. This implies, there is still abundant room for further progress.

Bibliography

[1] Soriano JB, Kendrick P, Paulson K, Gupta V, and Vos T, *Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017*, The Lancet Respiratory Medicine Journal (2020), 8: 585-96.

[2] *The burden of lung disease*, [Online]. Available on: https://www.erswhitebook.org/chapters/the-burden-of-lung-disease/. Access Date: May 23, 09:33.

[3] *Wold Lung Day 2019: Healthy Lungs For All,* [Online]. Available on: https://goldcopd.org/world-lung-day-2019-healthy-lungs-for-all/. Access Date: May 23, 10:12.

[4] UNICEF, *Fighting for Breath: A Call to Action to Stop Children Dying from Pneumonia,* [Online]. Available on: https://data.unicef.org/resources/fighting-for-breath-a-call-to-action-to-stop-children-dying-from-pneumonia/. Access Date: May 24, 11:28.

[5] [Online]. Available on: www.worldlifeexpectancy.com, Access Date May 23.

[6] Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F, *International Agency for Research on Cancer*, Global Cancer Observatory: Cancer Today. Lyon, France, (2018), [Online]. Available on: https://gco.iarc.fr/today. Access Date: May 24, 13:11.

[7] World Health Organization, *Air Pollution*, [Online]. Available on: https://www.who.int/health-topics/air-pollution/. Access Date: May 24, 11:35.

[8] J. Ma, Y. Song, X. Tian, Y. Hua, R. Zhang, and J. Wu, *Survey on deep learning for pulmonary medical imaging*, Front. Med., vol. 14, no. 4, pp. 450–469, May 22, doi: 10.1007/s11684-019-0726-4.

[9] J. Hodler, R. A. Kubik-Huch, and G. K. von Schulthess, Eds., *Diseases of the Chest, Breast, Heart and Vessels 2019-2022: Diagnostic and Interventional Imaging*, Springer International Publishing, 2019.

[10] N. Maskell, *Pocket Tutor Chest X-Ray Interpretation*, London: JP Medical Ltd, (2012).

[11] J. S. Klein and M. L. Rosado-de-Christenson, *A Systematic Approach to Chest Radiographic Analysis*, in Diseases of the Chest, Breast, Heart and Vessels 2019-2022, J. Hodler, R. A. Kubik-Huch, and G. K. von Schulthess, Eds. Cham: Springer International Publishing, 2019, pp. 1–16.

[12] Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., McDonald, C. J. *Preparing a collection of radiology examinations for distribution and retrieval*, (2016) , Journal of the American Medical Informatics Association, 23: (2), 304–310 (p. 2, 3, 10, 16, 19, 22, 63, 88, 94, 107, 121)

[13] J. Yanase and E. Triantaphyllou, *A systematic survey of computer-aided diagnosis in medicine: Past and present developments*, Expert Syst. Appl., vol. 138, p. 112821, May 204, doi: 10.1016/j.eswa.2019.112821.

[14] A. K. Bharodiya and A. M. Gonsai, *An improved edge detection algorithm for X-Ray images based on the statistical range*, Heliyon, vol. 5, no. 10, p. e02743, May 23, doi: 10.1016/j.heliyon.2019.e02743.

[15] S. Akcay and T. Breckon, *Towards Automatic Threat Detection: A Survey of Advances of Deep Learning within X-ray Security Imaging*, ArXiv200101293 Cs, Jan. 2020, Accessed: May 24. [Online]. Available: http://arxiv.org/abs/2001.01293.

[16] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, *Key challenges for delivering clinical impact with artificial intelligence*, BMC Med., vol. 17, no. 1, p. 195, May 23, doi: 10.1186/s12916-019-1426-2.

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, *ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases*, p. 10, May 2017.

[18] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, *Learning to diagnose from scratch by exploiting dependencies among labels*, ArXiv171010501 Cs, Feb. 2018, Accessed: May,20. [Online]. Available: http://arxiv.org/abs/1710.10501.

[19] Y. Xia, *ChestNet: A Deep Neural Network for Classification of Thoracic Diseases on Chest Radiography*, p. 8, Jul. 2018.

[20] P. Kumar, M. Grewal, and M. M. Srivastava, *Boosted Cascaded Convnets for Multilabel Classification of Thoracic Diseases in Chest Radiographs*, ArXiv171108760 Cs, Nov. 2017, Accessed: May, 20, [Online]. Available: http://arxiv.org/abs/1711.08760.

[21] S. Guendel et al., *Learning to recognize Abnormalities in Chest X-Rays with Location-Aware Dense Networks*, ArXiv180304565 Cs, Mar. 2018, Accessed: May, 20. [Online]. Available: http://arxiv.org/abs/1803.04565.

[22] I. Sirazitdinov, M. Kholiavchenko, R. Kuleev, and B. Ibragimov, *Data Augmentation for Chest Pathologies Classification*, in 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, Apr. 2019, pp. 1216–1219, doi: 10.1109/ISBI.2019.8759573.

[23] Y. Ma, Q. Zhou, X. Chen, H. Lu, and Y. Zhao, *Multi-attention Network for Thoracic Disease Classification and Localization*, in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, May 2019, pp. 1378–1382, doi: 10.1109/ICASSP.2019.8682952.

[24] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, *Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification*, Sci. Rep., vol. 9, no. 1, p. 6381, Dec. 2019, doi: 10.1038/s41598-019-42294-8.

[25] Philips Healthcare, *SkyPlate detector technical data*, (2020), [Online]. Available on: https :/ /www. philips.de/healthcare/product/HCNOCTN343/skyplate-detektor-mobile-wlan-detektoren-24x3035x43/technische-daten, Access date : February 2022.

[26] He, K., Zhang, X., Ren, S., Sun, J., *Deep residual learning for image recognition*, (2015), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778 (see pp. 12, 21, 57, 58, 62, 68)

[27] *Cluster Analysis: Basic Concepts and Algorithms*, [Online]. Available on: https://www-users.cse.umn.edu. Access date: April 2022.

[28] Embrechts, M. and Gatti, Christopher and Linton, Jonathan and Roysam, Badrinath, (2013) *Hierarchical Clustering for Large Data Sets*, 10.1007/978-3-642-28696-4_8.

[29] *17 types of similarity and dissimilarity measures used in data science*, [Online]. Available on: https://towardsdatascience.com/17-types-of-similarity-and-dissimilarity-measures-used-in-data-science-3eb914d2681. Access date: March 2022.

[30] Day, W.H.E. and Edelsbrunner, H. *Efficient algorithms for agglomerative hierarchical clustering methods,* J. Classif., 1, pp. 7-24 (1984).

[31] Thomas M. Cover, Joy A. Thomas, *Elements of Information Theory*, 2^{nd} Edition, page 14, (2006).

[32] Serafin Moral, Andres Cano, Manuel Gomez-Olmedo, *Computation of Kullback–Leibler Divergence in Bayesian Networks*, Entropy, (2021), 23,1122. Available on: https://doi.org/10.3390/e23091122

[33] Frank Nielsen, *On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means*, (2019), Entropy 2019, 21, 485.

[34] Jianhua Lin, *Divergence Measures Based on the Shannon Entropy*, IEEE Transaction on Information Theory, vol 37, no.I, JANUARY (1991)

[35] F. Gomez-Lopera, J. Martinez-Aroza, A. M. Robles Perez, R. Roman-Roldan, *An analysis of edge detection by using the Jensen-Shannon divergence,* Journal of Mathematical Imaging and Vision, vol. 13, pp. 35–56, (2000).

[36] S. Mairhofer, S. Zappala, S. R. Tracy, C. Sturrock, M. Bennett, S. J. Mooney, T. Pridmore, *RooTrak : Automated recovery of three-dimensional plant root architecture in soil from X-ray microcomputed tomography images using visual tracking,* Plant Physiology, vol. 158, pp. 561–569, (2012.)

[37] J. Bassak, *Detection of neural activities in fMRI using Jensen-Shannon divergence,* International Journal of Biometrics and Bioinformatics, vol. 6, pp. 113–122, (2012).

[38] A. Ben Hamza, H. Krim, *Image registration and segmentation by maximizing the Jensen-Renyi divergence,* Lecture Notes in Computer Science : Energy Minimization Methods in Computer Vision and Pattern Recognition (A. Rangarajan, ed.), vol. LNCS 2683, pp. 147–163, Berlin : Springer, 2003.

[39] N. Gillard, E. Belin, D. Rousseau, F. Chapeau-Blondeau, *Divergence informationnelle de Jensen-Shannon appliquée à la segmentation d'images,* 26ème Colloque GRETSI sur le Traitement du Signal et des Images, Juan-les-Pins, France, 5-8 sept. (2017)

[40] Fangyan Nie, Jianqi Li, *Image Threshol Segmentation with Jensen-Shannon Divergence and Its Application*, IAENG International Journal of Computer Science, 49:1, IJCS, 49, 1, 21, (2022).

[41] Stanford ML Group, *CheXphoto Dataset*, [Online]. Available on: https://stanfordmlgroup.github.io/competitions/chexphoto/. Access date: November 2021.

[42] Python, [Online]. Available on: https://www.python.org/.

[43] Strichartz, Robert S. (2000), *The Way of Analysis*, Jones & Bartlett Learning, p. 357, ISBN 978-0-7637-1497-0

[44] Joe Briet, Peter Harremoes, *Properties of Classical and Quantum Jsensen-Shannon Divergence*, 9.70.Cf, 03.67.-a. (2009), [Online]. Available on: http://arxiv.org/abs/0806.4472v4.

[45] Frank Kruger, *Activity, Context, and Plan Recognition with Computational Causal Behaviour Models*, (2016), [Online]. Available on: https://www.researchgate.net/publication/314116591

[46] Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*, (2012), page 877.