M. Tech (Software Engineering) Rohan Tomar 2022

# CLUSTERING: ANALYSIS OF ALGORITHMS AND APPLICATIONS

A DISSERTATION

SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE

OF

MASTER OF TECHNOLOGY

IN

**SOFTWARE ENGINEERING**

Submitted by:

**Rohan Tomar (2K20/SWE/19)**

Under the supervision of

DR. ABHILASHA SHARMA

**DEPARTMENT OF SOFTWARE ENGINEERING**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of Engineering)

Bawana Road, Delhi-110042

**MAY 2022**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi- 110042

## CANDIDATE'S DECLARATION

I, Rohan Tomar (2K20/SWE/19) of M.Tech (Software Engineering), at this moment declare that the Project Dissertation titled "Clustering: Analysis of algorithms and applications" submitted by me to the Department of Software Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation.

Place: Delhi

Date: 30 | 05 | 2022

Rohan

ROHAN TOMAR

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi- 110042

# CERTIFICATE

I, Rohan Tomar (2K20/SWE/19) hereby certify that the Project Dissertation titled " Clustering: Analysis of algorithms and applications" submitted by me in the Department of Software Engineering, Delhi Technological University, Delhi, in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this university or elsewhere.

Place: Delhi

Date: May 30, 2022

Abhilasha Sharma
30/5/2022

DR. ABHILASHA SHARMA

SUPERVISOR

Assistant Professor,

Department of Software Engineering,

Delhi Technological University,

(Formerly Delhi College of Engineering),

Shahbad Daulatpur,

Main Bawana Road,

Delhi- 110042.

# ABSTRACT

With the rise of the application of Machine learning in academia and industrial sector, clustering has become an important field of study. Clustering has been extensively used in studies involving unlabeled data, image processing and unsupervised learning. We have taken up the concept of clustering for our study and we have discussed the performance of two very popular and well used clustering algorithms and the application of clustering. The application of clustering has been discussed in the context of COVID 19 and medical field. In the First part, we compare K-Means and BIRCH Clustering algorithms on multiple datasets, and derive our results. After considering those results, we would move to discuss the application part. As we know, the world is witnessing an unprecedented catastrophe as a result of the COVID-19 epidemic, which has spread to approximately 216 nations and territories throughout the globe. A COVID-19 infection may progress to pneumonia, which may be diagnosed by CXR (Chest X-Ray) examination and should be treated as soon as possible after diagnosis. This part would be intended to examine the use of artificial intelligence in speedy & accurate diagnosis of COVID-19 pneumonia utilizing digital CXR pictures. In this research, we use a machine learning (ML) method i.e. SVM (Support Vector Machine) classification technique. SVM was used in the development of the model. The purpose of this research has been to use clustering, image processing, image segmentation, and feature extraction in fast or accurate identification of COVID19 chest X-ray or CT images. We assessed the performance of ML techniques on chest X-ray pictures as well as CT scans to COVID-19 diagnosis in this work. The model's performance was assessed using relevant classification measures, such as accuracy, precision, recall, & F1 score, among others.

# ACKNOWLEDGMENT

I would like to express my appreciation to all those who helped to constitute and complete this project .This work would not have been possible without the support of Dr. Abhilasha Sharma, who have been supportive of my career goals and who worked actively to provide me with the protected academic time to pursue those goals. As my teacher and mentor, she has taught me more than I could ever give her credit for here. I am grateful to all of those with whom I have had the pleasure to work during this and other related projects.

**ROHAN TOMAR**

# CONTENTS

# LIST OF FIGURE(S)

# LIST OF TABLE(S)

# LIST OF ABBREVIATION(S)

| **Abbreviation(s)** | **Full Forms** |
|---|---|
| WHO | World Health Organization |
| COVID-19 | Coronavirus Disease-19 |
| CXR | Chest X-Ray |
| CT | Computed Tomography |
| CAD | Computer-Aided Diagnosis |
| SARS-CoV | Severe Acute Respiratory Syndrome Coronavirus |
| RT-PCR | Reverse Transcription Polymerase Chain Reaction |
| AI | Artificial Intelligence |
| ML | Machine Learning |
| DL | Deep Learning |
| PCA | Principal Component Analysis |
| KMC | K-Means Clustering |
| NBC | Naïve Bayesian Classifier |
| DT | Decision Tree |
| KNN | K-Nearest Neighbors |
| ANN | Artificial Neural Network |
| SVM | Support Vector Machines |
| RF | Random Forest |
| LR | Logistic Regression |
| AdaBoost | Adaptive Boosting |
| AEs | Autoencoder |
| RBM | Restricted Boltzmann Machine |
| CNN | Convolutional Neural Networks |
| RNN | Recurrent Neural Networks |
| DBN | Deep Belief Networks |
| FE | Feature Extraction |

# CHAPTER 1

# INTRODUCTION

## 1.1 OVERVIEW

Clustering has been extensively used in studies involving unlabeled data, image processing and unsupervised learning. The first purpose of this study is to discuss various applications of clustering, and take up two clustering algorithms of K means and Balanced Iterative Reducing and Clustering through Hierarchy Clustering (BIRCH), as these algorithms have been applied to a wide array of studies across different domains, and document a comparative study between these two. K Means is a technique of Partitioning based clustering, whereas BIRCH clustering is a technique of Hierarchical Method of clustering. The methodology is based on comparing BIRCH and K-Means Clustering algorithms, using a total of five datasets. The pre-processing of dataset is described in the later parts of the report; Encoding is performed on different attributes of these datasets and dimensionality reduction using principle component analysis is performed. In this study, validation of clustering performance using the Internal Validation is done, and here silhouette index is used. The Silhouette index takes into account the mean distance between the clusters. The results are tested on 2 to 100 clusters and deduction of optimal quantity of clusters is done as well. On every dataset it is observed that K-Means outperforms BIRCH clustering method. Thus we come to the conclusion K-Means clustering algorithm proves out to be better for clustering, for our considered datasets. So this result would be the basis for the next part of our two fold study, that is, explain and implement K-Means (as it is the better performing clustering algorithm) in the analysis and detection of COVID 19 disease which would explain the application of clustering in our study.

Clustering is a technique using which data is grouped into groups or clusters of alike data, and each cluster would belong to a class or label. This division of similar data points into classes have several applications in the field of data mining and machine learning (ML). Many published studies have used concepts of clustering in processing their data before the deduction of results. Clustering has found its application is many fields of studies, like customer analysis. This helps businesses identify their different customer bases having different characteristics of consumption. Similarly, recommendation systems also utilize this technique. One major area clustering is employed is image processing. Clustering is increasingly being employed for image segmentation. Image Segmentation (IS) is a technique via which division of an image into different parts or ‘segments’ is done, after which the image turns more meaningful, or often easier to process by a ML algorithm. Clustering is of many types and numerous algorithms can be used do perform it. Figure 1.1 shows the different types of clustering.

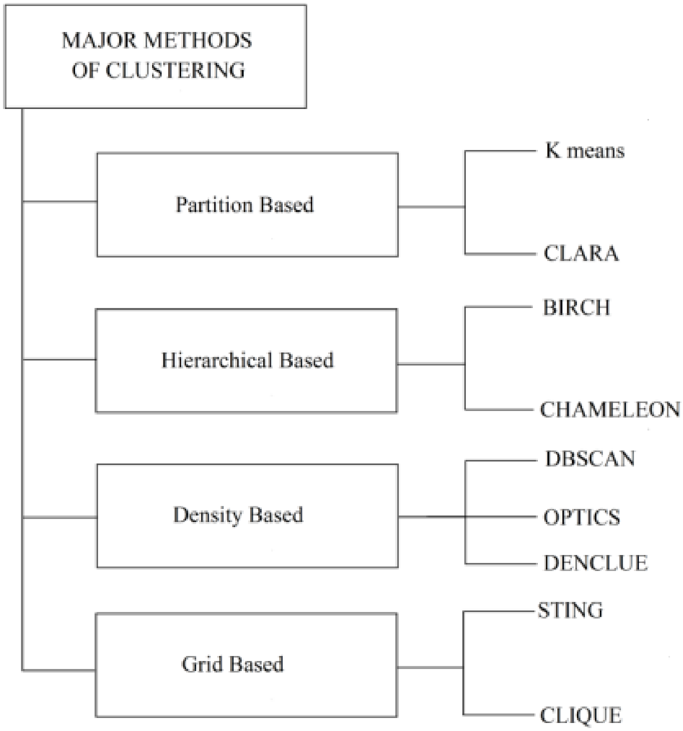


Figure 1.1: Types of Clustering.

COVID-19 is a lethal illness caused by a coronavirus that has just recently been identified. SARS-CoV, a coronavirus that entered the human body for the first time in December 2019, may transmit primarily among people via droplets created by infected patients when they talk, cough, or sneeze, as well as through the atmosphere (Nur-a-alam et al., 2021) [1]. Because the droplets are too weighty to travel long distances, they are unable to spread from person to person while coming into close touch with one another (Holshue et al., 2020) [2]. Even though the precise period is not yet established, recent research estimates that COVID-19 may continue functioning in the air for up to 3 hours, on copper for 4 hours, as well as on plastic as well as corrosion resistance for up to 72 hours, depending on the material. The specific responses to these issues, on the other hand, are all still up in the air within the broader health academic researchers but are now under examination. Many individuals may not notice any severe symptoms, although the majority of patients had fever & cough as primary symptoms. It is common to have other secondary symptom such as body pains, sore throat, as well as a headache as well.

COVID-19 illness is now spreading at an alarming rate, owing to a lack of rapid diagnosis technologies. In 2020, a large number of individuals will die as a result of this illness all over the globe. The respiratory systems, as well as lungs, are the primary sites where the virus may spread quickly and efficiently. This results in an inflammatory response and the formation of fluid-filled air sacs that may then be discharged. The mechanism is responsible for establishing a stumbling block in the intake of oxygen. The rapid and precise diagnosis of the virus is a big issue for physicians and other health care workers all around the globe, but it is essential if the fatality rate generated by this virus is to be reduced.

People were already suffering from a variety of different ailments as a result of global environmental issues, as well as the COVID-19 will have an immense influence on this situation. In its current state, the virus has migrated to nearly every country in the region. As a result, the countries with the highest revealed that the current COVID19 cases are America, South-East Asia, and Europe, respectively (Figure 1.2). As of the 7th of January, 2021, the World Health Organization (WHO) has recorded more than 85,929,428 confirmed cases of the virus and more than 1,876,100 fatalities as a consequence of the epidemic. To diagnose viral infections and isolate those who have been infected, further research into efficient screening procedures is needed at this time, according to the World Health Organization. International efforts by healthcare professionals as well as scientists to strengthen their treatment plans or test capacity via the use of multifunctional testing are underway in many nations across the globe to both boundary spread of virus & protect themselves from the fatal infection.

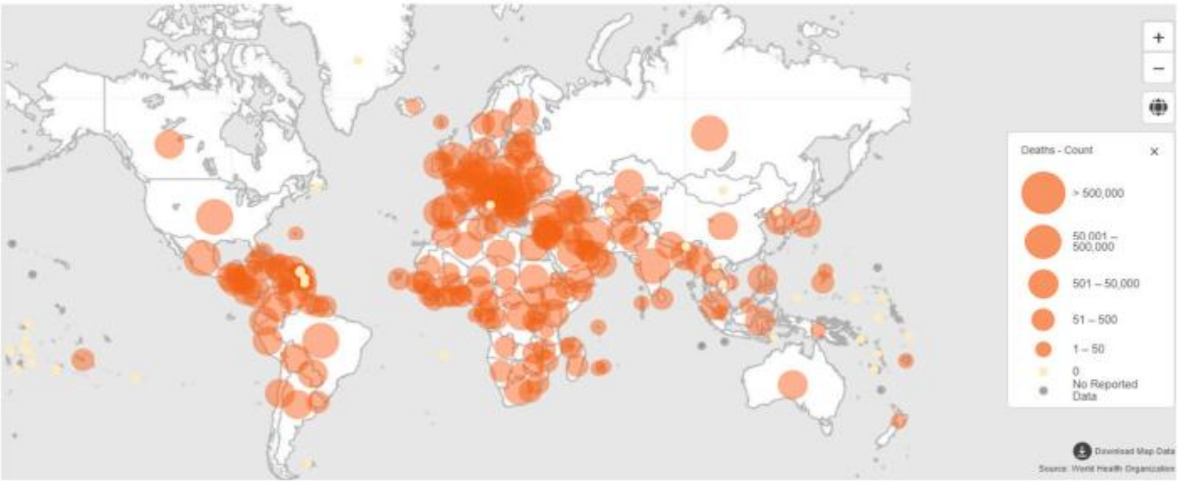


Figure 1.2: Map for COVID deaths according to WHO on 7 January 2021.

The epidemic of COVID-19 started with the first diagnosis of an unknown bacterial meningitis in late 2019 in Wuhan, China, and spread about world as a pandemic in following months. COVID-19 was discovered to be produced through a new Coronavirus (called severe acute respiratory syndrome coronavirus 2), which was discovered by ribonucleic acid sequence of respiratory samples collected from patients. Patients with COVID-19 appear with symptoms that are comparable to those of other viral infections, including such influenza as well as other coronaviruses including such acute respiratory syndrome, as well as other viral disorders. It is difficult to diagnose since the symptoms are vague. They include fever, cough, weariness, Dyspnea (Shortness of Breath), diarrhoea, & perhaps even Anosmia (Loss of smell). The radiographic findings are generic, and they may be noticed in individuals who are suffering from various viral diseases, medication responses, or aspiration as well (Zhang et al., 2021) [3].

The similarity in clinical presentation with different responses and disorders presents difficulties in establishing a clinical diagnosis in certain cases. Modern reference standard methods for identifying patients with COVID-19 infection include RT-PCR. In complement to the RT-PCR test, computed tomography (CT) has been routinely employed in China, but rarely in other countries, to give an alternative way of COVID-19 diagnosis as well as treatment response surveillance. Health care professional groups, on the other hand, do not suggest Imaging techniques as a standard diagnostic imaging technique for patients with COVID-19 because to considerations about CT facility contaminants including vulnerability to health care employees(Zhang et al., 2021) [3].

Major medical organisations, on the other hand, urge the usage of chest radiography as part of work-up for those who are suspected of having COVID-19 because of the particular benefits it provides: Just about all the clinics, emergency departments, urgent care centres, and hospitals, including both agricultural and non - agricultural medical institutions, are prepared with fixed & mobile radiography systems, which may be found in almost any setting. Because these units may be readily shielded from exposure & sanitised after use, it is possible to employ them immediately in a controlled clinical setting without the need to shift the patients. In present radiologic practise, the poor detection accuracy of chest radiography in the identification of COVID-19 provides a significant obstacle to the adoption of this technique. According to the findings of a recent research, chest radiography has low sensitivity for the identification of COVID-19. Approximately one-third of radiologists are viewing COVID-19–brought pneumonia for first time, & as a result, they must interpret more pictures in order to get familiar with both the common and unusual imaging characteristics of this illness (Zhang et al., 2021) [3].

With technical breakthroughs in extent of ML, X-ray pictures of the chest may now be utilised to diagnose COVID-19 using X-ray images of the chest. In the field of ML, DL has emerged as most important techniques. Deep learning focuses on collecting characteristics from pictures and classifying them, which may be used in a variety of applications such as object detection or, in medical circumstances, task categorization. When it comes to applying artificial intelligence to data mining, analysis, and recognition, ML and DL have established themselves as well-established fields in the field. COVID-19 may now be discovered, measured, and tracked thanks to recent advancements in AI, making it simpler to separate people who have been infected in order to get speedier treatment for their illness. Chest x-rays are used to screen the patient's lungs for infections that may be present. It is a more efficient, simpler, less expensive, and less dangerous means of testing patients, and as such, it must be adopted. All nations must employ chest X-rays immediately in light of the rising death rates, which may be traced back to the implementation of COVID-19. However, despite the fact that this technology advance appears to be beneficial, the images of multiple kinds of pneumonia are equivalent as well as overlap with pictures of other infectious as well as parasitic lung diseases, making it difficult to identify among COVID-19 and also other viral cases of pneumonia (Of, 2020) [4].

Investigation in Artificial Intelligence (AI), particularly in Machine as well as Deep Learning methods, has shown that these algorithms, when applied to medical pictures, perform very well. When it comes to quickly and tirelessly learning to discriminate COVID-19 pneumonia from other forms of pneumonia using chest radiographs, machine learning technologies, especially deep learning, offer distinct benefits. It was our goal in this research to grow & estimate a classification algorithm for distinguishing COVID-19 Analysis from Chest X-ray Photographs from other causes of malformations at computed tomography from other causes of anomalies at computed tomography, as well as to test its effectiveness against thoracic radiographers (Rishabh Raj, 2020) [5].

## 1.2 PERFORMANCE COMPARISON OF ALGORITHMS

We have taken up two algorithms for the comparative evaluation for the part 1 of the study. The two clustering algorithms are K Means Algorithm and the BIRCH Clustering algorithm. In the First part, we will compare these two on five datasets, and the performance would be measure on the silhouette coefficient. This part deals with the comparison of the algorithms.

Clustering, as a concept of data mining has been used in various real life examples for solving problems. In important application like medical imaging, K Means is a popular algorithm which is often used and gives good results. Studies have been conducted where K Means has been used with the combination of another algorithm fuzzy c means, to develop segmentation method for human brain magnetic resonance imaging (MRI) images [6]. A previous study has also used K means algorithm for the image segmentation of medical images to detect breast cancer [7]. K means has been used in the recognition of Parkinson’s disease, where K means algorithm has been used to cluster the fuzzified pixel information & K means was also used in the segmentation of MRI images for the same study [8]. K means is well performing, and it has been used in other domains too. These examples involve studies like prediction of churn in banking system, where K means algorithm has been applied along with other clustering techniques [9]. BIRCH Algorithm has also been used for various applications and has proved out to be a good performing algorithm for clustering data. An article has developed an improved version of BIRCH algorithm called Dynamic BIRCH (D BIRCH), and used to cluster 10,000 data points from the electricity generated from a thermal power plant [10]. BIRCH algorithm is also utilized in a study which aims to identify the malware families via clustering [11]. This study shows that BIRCH is a good performing algorithm for malware family identification. A previous study also uses BIRCH algorithm for the realization of different scenarios relating to wind power outputs [12].

We aim to do a comparative study based on these two algorithms and since these two clustering methods are very popular, well performing and have been tested in various case studies across different domains, Partitioning method clustering i.e. K-Means clustering, and Hierarchical Method i.e. BIRCH are taken up for the comparative evaluation in our article. Five datasets are considered for the comparative evaluation on these two algorithms. These two different algorithms have selected because primarily they belong to a different type of clustering method as illustrated in the Figure 1.2, and since both have given good results in the past, we also want to conduct a study which would help future researchers what algorithm to select for their data. These five datasets have a detailed description later on in the study. Evaluation of the clustering results is performed with the internal clustering validation metric called the silhouette coefficient. The methodology of the study is also described later on in the study.

## 1.3 CLUSTERING VALIDATION & RELATED METRICS

In this study, validation is done on grounds of a validation metric called the Silhouette Coefficient. The measure used to evaluate the clustering results is silhouette index. Every cluster is basically a silhouette, that displays which objects lie within its cluster, or which ones lie between clusters [13]. This measure has the index of [-1, 1]. The Silhouette value s(i) of a single data point Xi is calculated as shown in equation (1.1):

$$s(i) = (B(i) - A(i)) / \max\{A(i), B(i)\} \qquad \ldots\ldots(1.1)\ [14]$$

Where,

‘A’ stands for mean intra cluster distance, which indicates how compact the cluster that ‘i’ belongs to is. And ‘B’ stands for the mean nearest cluster distance, which indicates how far apart ‘i’ is from other clusters

Some other clustering validation techniques are also there. Calinski Harabasz index is on clustering validation technique which validates the clustering results based on average within cluster sum of squares. Index I is another coefficient which validates based on maximum distance between cluster centers and evaluates compactness based on the sum of distances between objects and their cluster center.

## 1.4 COVID-19 DIAGNOSIS IN CHEST X-RAYS IMAGES

After the completion of the comparative evaluation of clustering algorithms, we move on to the application of clustering in detection of COVID 19. We are using the clustering concepts the second part where COVID 19 is being detected with the help of a ML model. Many different technologies have been used in this part, like Principle component analysis (PCA), Image preprocessing using Unsharp filter method, but the clustering is being used to segment the images of the Chest X ray images before feeding them into a ML model. Segmentation is basically dividing the image into different regions so that the accuracy of ML model is increased.

COVID-19 pandemic, which was triggered by the infection of humans with the SARS-CoV, has continued to have catastrophic impact on health & well-being of worldwide population. A vital stage in battle against COVID-19 is adequate screening of diseased individuals, so that those who have been affected may get timely treatment and care, as well as be isolated in order to prevent the virus from spreading further across the population. The RT-PCR test is the major screening method for detecting COVID-19 patients. Additionally, the specificity of RT-PCR testing is quite diverse but has not been reported in a fair and clear fashion to far, with early findings in China showing suggesting it has rather low specificity. Also found were drastically varied positive rates based on how the information was received, as well as a diminishing positive rate as time passed after the onset of the illnesses(Das et al., 2021) [15].

A second screening approach for COVID-19 that has been used is radiography examination, where radiologists perform and evaluate chest radiography imaging to check for visual ciphers related with the SARS-CoV virus infection. Several early investigations have shown that individuals with COVID-19 infection had abnormalities in their chest radiographs, leading some to argue that radiography evaluation might be utilised as the main screening method for COVID-19 in epidemic regions. COVID-19 positive subjects in Huang et a research’s had bilateral radiographic abnormalities in CXR pictures, while Guan et a study’s had radiographic abnormalities like ground-glass opacity, bilateral abnormalities, and interstitial abnormalities in CXR and CT images. There has been a lot of talk about using CT imaging for COVID-19 screening because it provides better image detail in merger and acquisition, but there are several major benefits to using CXR imaging, especially in areas with limited resources and those that have been strictly impacted with global COVID-19 pandemic(Wang et al., 2020) [16]:

- **Rapid triaging:** When combined with viral testing (that also takes time), The use of CXR imaging enables for the quick triaging of individuals who have been diagnosed with COVID-19. If viral test method is not an alternative available, this process can be done in comparison to viral diagnostics (which requires months as well) to help alleviate the rising numbers of patients, particularly in regions most severely impacted where ability has been managed to reach, or as a stand-alone procedure if viral test method is not an alternative. Even before patients suspected of COVID-19 arrive at outpatient department, CXR imagery could be immensely beneficial for quarantining in geographical locations where patients are recommended to stay in the house until the onset of innovative symptoms. It’s because aberrations are commonly noted at the beginning of the survey because once patients secretive of COVID-19 arrive at outpatient department.
- **Availability & accessibility:** CXR imaging is commonly inexpensive is approachable at several clinical sites as well as imaging facilities since it is considered standard instrumentation in most health services. CXR imaging, in example, is considerably more widely accessible than CT imaging, which is particularly important in poor nations wherever CT scanners are prohibitively expensive owing to high operational and maintenance expenses.
- **Portability:** By having transportable CXR equipment, imaging may be done inside an isolation room, minimising the danger of COVID-19 transmission both during transfer to provides a detailed including such CT scanners and within the rooms containing the fixed imaging systems by a factor of many orders of magnitude.

## 1.5 COMPUTER-AIDED DETECTION (CAD)

For cases like lung cancer, computer-aided diagnosis (CAD) is most beneficial when it is able to portray lung cancer in people who are at elevated risk but do not exhibit disease-related symptoms, indicating that the tumour has been found at an early stage, when it is associated with a better prognosis. In addition to being the most prevalent examination technique in medical practise, chest radiography is also a valuable clinical tool in the identification of diseases. The automated identification of chest illness using chest radiography has therefore emerged as one of the most hotly debated areas in medical imaging research today. The research undertakes a complete review of computer-aided detection (CAD) systems on the basis of clinical applications, with a particular emphasis on the AI skill utilised in chest radiography.

CAD is a technique used in area of Radiology that provides vital information for surgical purposes. Because of its significance, numerous computer vision approaches are processed in order to retrieve valuable information from pictures obtained from imaging technologies e.g. X-ray, MRI, &CT scans, among others. Figure 1.3 shows CAD.

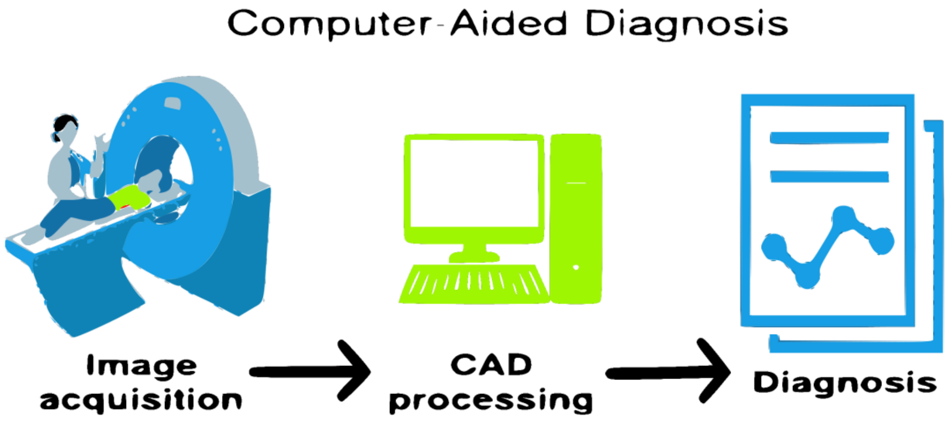


Figure 1.3: Computer-Aided Detection (CAD) System

Chest radiography (also known as CXR) is a medical imaging device diagnostic skill that is both affordable & simple to use. Currently, it is the most often used diagnosing tool in medical practise, and it evaluation of lung illness (Qin et al., 2018) [17]. Chest X-rays are performed by radiologists who have received extensive training in order to identify ailments including such pneumonia, TB, interstitial lung disease, as well as early lung cancer.

## 1.6 X-RAYS

X-rays are an essential medical tool for physicians to use in their practise. X-rays are ionised kinds of radiation that use rays to capture images of the subject. Doctors discovered that they were unable to get a comprehensive glimpse of the patient's physique. When it comes to seeing comprehensive information, alternative equipment such as MRI or CT scans is needed, which is much more costly. X-ray scans do not provide any medical information about organs or tissues; they solely provide an image of the bones. MRI and CT scans may provide more detailed images of the bones than ordinary X-rays. Unlike an X-ray, which provides a 2-D picture of the bone structure, a CT scan has the capability of providing a 3-D image of the bone structure However, MRI and CT scans are more expensive, putting them out of reach for most patients. As a result, digital x-ray technology is a solution for x-rays that exhibit a three-dimensional digital picture structure. Numerous variables contribute to the poor quality of an X-ray picture, both externally and internally. For illustration, for the external element, insufficient equipment, operator error, abnormality of the patient, while others are all factors that contribute to a poor image quality. For example, a lack of detail, poor contrast, and brightness in X-ray pictures are all possible outcomes of this situation. As a result, we need to increase the excellence of X-ray picture such that it is superior to prior image. Histogram equalisation may be used to allow for consistent lighting, adjusting grey levels to minimise noise, and utilising High-pass filters to clarify features, among other methods of improving X-ray images, among others. Since its invention, X-ray has found widespread use in scientific and medical disciplines.(Aziz et al., 2017) [18].

When it comes to assessing and identifying lung cancer, the chest X-ray picture requires the aid of developing CAD approaches. Typically, lung segmentation including nodule detection are carried out with the use of a computer-aided detection scheme, and nodule segmentation and diagnosis are carried out for the aid of a computer-aided detection technique. Cancer detection & detection in chest X-ray pictures is based on lung segmentation, which is the foundation of all other processes. Lung cancer detection and diagnosis, which is based on correct specificity and sensitivity of lung segmentation, resulting in patients receiving early treatment as well as extending the life of cancer patients(*Image Pre-Processing Techniques for X-Ray Medical Images : A Survey*, 2021) [19].

## 1.7 CHEST X-RAYS

Since its inception, chest X-ray imagery has been a core component of radiological imaging, and so it remains the most commonly performed radiological standardized test in the globe, to developed nations reporting an estimate of 238 enlarged chest X-ray captured images per 1000 of demography per year in industrialized nations. A total of 129 million CXR pictures were captured in the United States alone in 2006, according to current estimates. Because of the low radiation dosage and cost-effectiveness of CXR pictures, as well as their acceptable sensitivity to a broad range of diseases, there has been an increase in the demand for & affordability of CXR images. CXR is frequently imaging study performed, and it continues to be important in the screening, diagnosis, and therapy of a wide variety of disorders.

There are three primary kinds of chest X-rays that may be distinguished based on the location as well as direction of patient in relation to X-ray source & detector panel: Posteroanterior, anteroposterior, and lateral are all terms used to describe the position of a body part. In radiology, the frontal views are known to it as the posteroanterior (PA) and anteroposterior (AP) views, accordingly, since the X-ray source is positioned to the back or front of the person in both circumstances. The AP image is typically taken from people who are lying down, while the PA picture is commonly produced from patients who are standing up. Taking a lateral image in combination with a PA image is usual, and thus projection the X-ray through one side of the subject to another, most typically from bottom to top, is what is meant by lateral imaging. Figure 1.4 illustrates several examples of various picture kinds in more detail (Çallı et al., 2021) [20].

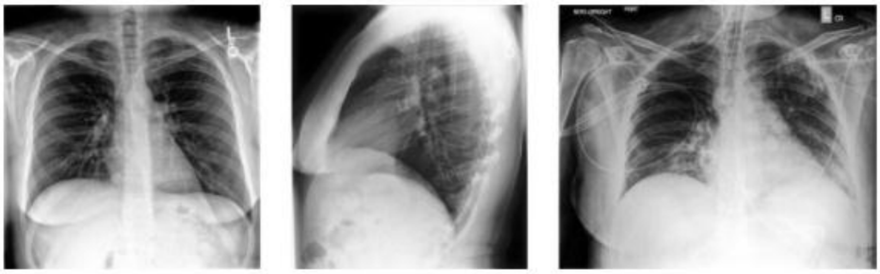


Figure 1.4:  Left: PA view frontal chest radiograph. Middle: lateral chest radiograph. Right: AP view chest radiograph.

Because of the superimposition of anatomical features along the projected plane, interpretation of chest radiograph might be difficult. As a consequence of this effect, it may be very hard to identify abnormalities in specific areas, to perceive minor or subtle anomalies, or to reliably discriminate between distinct clinical patterns on same image. These factors contribute to substantial inter-observer variability seen in examination of CXR pictures by radiologists in general (Çallı et al., 2021) [20].

## 1.8 MACHINE LEARNING

Since our application has a major part of ML which is used for classifying the images into different categories, it is important to discuss the ML part in this section.

Medical imaging has completely transformed the health-care business, allowing physicians and scientists to have a more in-depth understanding of human body than they ever had before. Medical pictures are also used to follow progression of an ongoing condition as well as to give patients with a more thorough level of treatment. No doubt, computed tomography provides a more comprehensive and better picture, but it is expensive and not readily accessible in underdeveloped regions due to the lack of infrastructure. A qualified radiologist can diagnose pneumonia, cardiovascular breakdown, lung sickness, and other conditions based on interpretation of CXR. The affordability, simplicity, and non-invasive nature of XR technology make it a popular choice. The procedure takes just a few minutes, and the findings are available within minutes of the procedure being completed. Even in underdeveloped areas, advanced digital radiography devices are readily accessible. Thus, chest radiographs are often performed only for screening purposes. The most difficult aspect of examining chest X-rays is necessity for presence of an experienced radiologist. The analysis of a chest X-ray is dependent on skill of radiologist since various bodily components overlap in a chest X-ray, which may conceal sick tissue. Consequently, more work is needed to build a computerised automated technique to aid radiologists in classifying one lung or both lungs with pneumonia and a normal lung from chest X-ray pictures, which is currently lacking. Recently developed AI skill provides novel potential in area of CAD systems to identify illnesses automatically from chest X-rays using ML. ML procedure learns from huge input data that has been labelled. (Parveen & Khan, 2020) [21].

An area of AI known as ML is a technique that tackles real-world issues by "giving learning capabilities to a computer without the need for extra programming".

The attempts to determine if computers might accumulate information in order to emulate the human brain resulted in the development of machine learning techniques. When Arthur Samuel created the first game-playing software for checkers in 1952, he was attempting to achieve enough abilities to defeat a world checker champion. This was the beginning of ML's journey. Later, in 1957, Frank Rosenblatt developed an electrical system that can learn how to solve complicated problems by emulating the process that occurs in the human brain, which he named the Rosenblatt Machine. The advancement of machine learning contributes to the increased use of computers in medicine (Alić et al., 2017) [22].

In the majority of situations of illness identification and diagnosis, the development of machine learning systems is seen as an effort to mimic the expertise of medical specialists in the identification of sickness. Because machine learning (ML) allows computer programmes to learn from data, creating a program to recognise similar themes including being able to decide on measure of academic, it does not have difficulty coping with the insufficiency of the medical database that is being utilised to train it. Classification the most well ML approach in medical applications because it relates to issues that arise in daily life, and it is also the methodology that is most often used in medical applications(Alić et al., 2017) [22] (Zhou et al., 2015) [23].

### 1.8.1 Types of Learning

A machine learning system learns from its previous experiences in order to increase the effectiveness of intelligent software components on a computer. Machine learning systems may be divided into two categories(Reddy & Babu, 2018)(Hormozi et al., 2012) [24][25]:

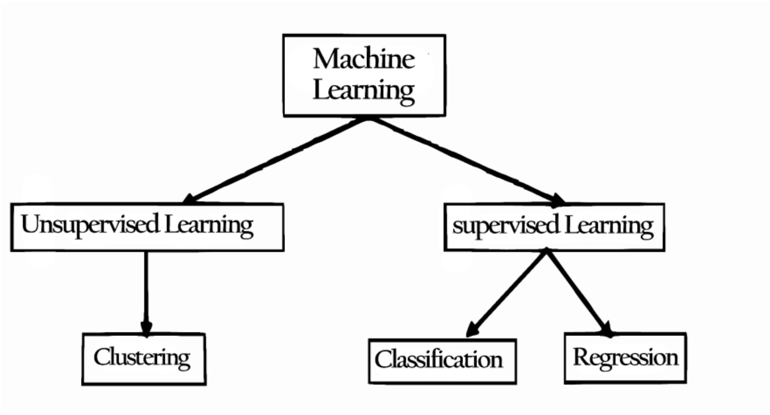


Figure 1.5: Types of Machine Learning

#### *1.8.1.1 Supervised Learning*

It is necessary to have training data that contains labelled data and data that has a producing value in order to do supervised learning. In the case when a collection of algorithms needs external help, the techniques are referred to as supervised machine learning algorithms. In these methods, input dataset is divided into 2 parts: training dataset as well as testing dataset. And the output variable that has to be predicted or categorised comes under the train dataset category as well. These results are achieved in such a manner that certain types of patterns from the training dataset are learnt by all of the methods. These are now performed to the test dataset in order to predict or sort the data (Meenakshi, 2020) [26].

#### *1.8.1.2 Unsupervised Learning*

Unsupervised learning methods are used in this case. Do not utilise a training set as well as attempt to discover patterns or organization in the data on your own. Unsupervised clustering issues may be tackled applying a diversity of methods. This technique is those that learn a minor no. of characteristics from data without the assistance of a human being or a computer. Whenever new data is fed into or introduced into these techniques, the class of the data is identified based on the previously learnt characteristics of the data. This sort of technique is mostly used for classification and feature reduction tasks in software. Clustering and dimensionality reduction approaches are two of the most used kinds of algorithms, and they are divided into two categories. (Meenakshi, 2020) [26].

- **K-Means Clustering:** KMC is a form of unsupervised learning technique that creates discrete clusters in the number 'K'. When the programme is started, it automatically forms clusters or groups. The objects that have the same characters are grouped in the same cluster using this technique approach. The centre of a certain cluster is essentially the mean of the individual clusters in terms of distance.
- **Principal Component Analysis:** By lowering the number of dimensions in the data, the PCA algorithm approach makes calculations simpler and much more rapid. PCA is an abbreviation for principal component analysis. We may use 2D data as an example to further comprehend this straightforward method. The data normally takes up two axes when displayed on a graph, but when PCA is applied

to the data, the data becomes one axis, allowing for simple and rapid calculations to be performed.

### 1.8.2 Machine Learning Technique

Different machine learning models are present to distinguish between chest X-ray pictures that were abnormal and those that were not. Providing a binary classification of the inclusion or exclusion of influenza on a CXR is goal of this module, as is selecting the most appropriate model for pneumonia forecasting.

#### *1.8.2.1 Naive Bayesian*

It is simple to construct a NB model since it does not need sophisticated iterative parameter estimates, making it especially effective for extremely big datasets. Due to its simple structure, the NBC often beats more advanced classification algorithms, and so it frequently performs shockingly well. It is commonly used because of its effectiveness. NB classification method is widely used classification techniques in the field of data mining. NBC is a fundamental probabilistic-based algorithm that predicts the likelihood of a class member belonging to that class (Han et al., 2012) [27]. If an attribute has an impact on a specific class that is also irrespective of the effects of other characteristics, this is referred to as conditional probability independent in the context of a NBC.

#### *1.8.2.2 Decision Tree*

A DT is a categorization model that is built in the shape of a hierarchy. DT learning, which is used in data mining as well as ML, is a probabilistic classifier that maps information about in an item to inferences about just the item's target value. It is a DT that is employed as a predictive model. Classification technique and regression trees are two more descriptive terms for these types of modelling techniques (Witten et al., 1999) [28]. The leaves of these tree constructions indicate categories, while the branches reflect the conjunctions of elements that result to the classifications represented by the leaves. An example of a decision tree is one that may be used in decision processes to graphically as well as clearly describe choices as well as decision making. In data mining, a decision tree explains data but does not make judgments; rather, the categorization tree that is produced may be used as an input for setting priorities.

#### *1.8.2.3 K Nearest Neighbours*

KNN (Aha et al., 1991) [29] is most straightforward classification techniques available in ML. It falls under the category of instance-based learning, sometimes known as lazy learning. Using this classification strategy, it is possible to take into consideration local approximation while deferring all calculation till the categorization process is complete. It preserves all of the recorded examples in the supplied dataset as well as classifies new cases similarity measures including such Euclidean distance. As a result, the K most comparable occurrences, also known as the neighbours, are managed by searching across the whole prepared set for yet another test data point, which is then controlled. Furthermore, the anticipated result is obtained by abrogating the yielding parameter for those K instances that are reliant on a majority vote of the neighbours in the greater portion of the neighbourhood.

#### *1.8.2.4 Artificial neural network (ANN)*

An ANN, sometimes known as a neural network (NN), is a mathematical proof or computer model that is inspired by the way biological and/or functional elements of biological NNs (e.g., the hippocampus). NN is made up of a no. of artificial neurons that are linked to one another and that processing information using a chosen because it provides to computing. In the majority of situations, an ANN is an adaptive integrative process in response to external or requires a structure that travels through the network during in the learning phase of the algorithm. Modern neural networks are statistical data modelling techniques that are non-linear in nature. They are often used in the modelling of complicated interactions between inputs and outputs, as well as in the discovery of patterns from the data (Alawnah & Sagahyroon, 2017) [30].

#### *1.8.2.5 Support vector machines (SVMs)*

Classification as well as regression analysis are both performed using SVMs, which are a group of similar supervised learning algorithms that examine data and detect patterns. SVMs (Keerthi et al., 2001) [31], are yet another prominent classification approach that is frequently employed in a variety of predictive analytics applications. Typically, SVM is used as a binary classifier, which means it is used to data from the following emphases in informational variable space according to their class, which may be either class 0 or class 1. An appropriate hyperplane is selected in vector machine that is a line that may participate in the variable space and is used to do this task. With the help of this hyperplane, the vector ML computation is able to determine the coefficients that result in the optimal detachment of the subclasses.

#### *1.8.2.6 Random forest (RF)*

RF classifier introduced by (Breiman, 2001) [32] is an ensemble machine learning approach that generates a prediction result by taking into consideration numerous learning algorithms at the same time. In DT building, RF is a strategy that uses bootstrap aggregation (bagging) in conjunction with random feature extraction to produce a collection of DTs that have controlled variance. Instead of a single DT, RF generates a huge number of DTs to anticipate the final output behaviour class of mobile phone users based on their cell phone log data, rather than a single decision tree. The over-fitting problem that may arise with a single DT, that was earlier resolved, is eliminated since it creates numerous decision trees for a particular dataset.

#### *1.8.2.7 Logistic regression (LR)*

LR (Cessie & Houwelingen, 1992) [33] is yet another common probabilistic-based statistical approach for solving classification difficulties that is utilised to address classification tasks. Most of the time, logistic regression is used to estimate probabilities by applying a logistic function that is also known as the sigmoid function. The hypothesis of logistic regression suggests that function should be limited to values between 0 and 1. Using a given dataset, this classifier determines the link between a categorical dependent variable as well as one or more independent factors. This is the dependant variable, which is the target class that we will forecast. The independent variables, on the other hand, are the qualities or contextual factors that we will use to forecast which class will be selected.

#### *1.8.2.8 Adaptive boosting (AdaBoost)*

This is a ML meta-algorithm developed by (Freund & Schapire, 1996) [34], which stands for adaptive boosting. This approach may also be used to make predictions in certain situations. Boosted classifiers are designed to integrate the output of other learning procedures, known as 'weak learners,' in order to build an operative classifier that can be used to get the final output of the boosted classifier method. This is referred to be an adaptive classifier in that it significantly improves the performance of the classifier; nevertheless, in certain circumstances, it may result in overfitting. This is sensitive to noisy data & outliers, and that may be used to improve performance. A lot of experts in the field of context-aware mobile services make use of the AdaBoost classifier for a variety of applications.

DL is a subset of ML, which is wide phrase that refers to methods for learning new things. DL has risen to prominence in recent years as approach of choice for IP jobs, and it has had a significant influence on area of medical imaging. DL is notoriously data-hungry, & indeed CXR research communal has stands to benefit from publishing of many large labelled databases in recent years, which were primarily made possible by the automatic data preprocessing of radiology reports, which allowed for generation of labels in the first place.

Deep learning is a new study topic in the fields of machine learning as well as pattern recognition that is rapidly gaining popularity. When it comes to categorization, deep learning is a term used to refer to machine learning approaches that employ supervised or unsupervised procedures to automatically learn classification tasks in deep structures. In deep learning, the notion of a human brain with various forms of representation, with basic characteristics at the lower levels and higher-level abstractions constructed on top of that, is referred to as multi-level representation. Humans organise their thoughts and concepts in a hierarchical fashion. Initially, human beings absorb basic notions, which they later combine to represent more complex concepts. When compared to a deep neural network, the human brain is made up of several layers of neurons that operate as feature detectors, recognising more abstract characteristics as the layers of neurons are increased in number. This method of describing information in a more abstract manner is simpler for computers to generalise than the previous one. The goal is to identify more abstract characteristics at the upper layers of the representations by employing neural networks, which are capable of quickly distinguishing between the numerous explanatory elements included in the data. Because of its state-of-the-art effectiveness in a wide range of domains such as object perception, voice recognition, computer vision, collaborative filtering, including natural language processing in recent years it has gained a great deal of interest. The amount of data being generated is growing at an exponential rate, and deep learning is becoming more important in offering big data predictive analytics solutions (W. Liu et al., 2017) [35].

### 1.8.3 Overview of Deep Learning Techniques

There has been a plethora of DL-based designs suggested. Deep architectures such as CNNs, AEs, RNNs, as well as DBNs are amongst the most often utilised deep architectures for detection of DR. Various architectural styles will be discussed in detail in the following sections (Asiri et al., 2019) [36].

#### *1.8.3.1 Auto encoder*

The standard definition of an auto encoder is a feed forward neural network with the goal of learning a compressed, distributed representation of a dataset. An auto-encoder is a three-layer neural network that has been taught to recreate its inputs by utilising them as the output of the network. Because it has to automatically learn which represent the variation in the data in order for it to be replicated, Using simply linear convolution layers, it can be demonstrated to be comparable to PCA and might even be utilised for dimensionality reduction. Once trained, the hidden layer activations are utilised as the learnt features, as well as the top layer may be removed from consideration altogether.

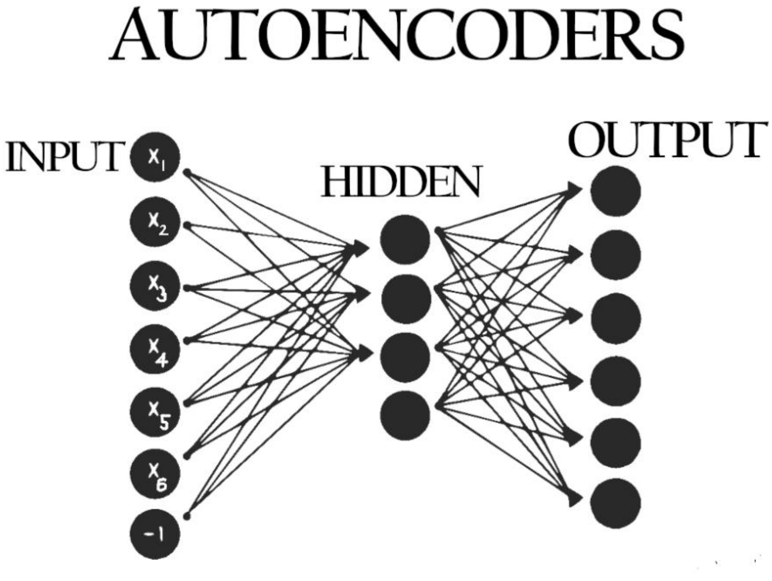


Figure 1.6: Auto encoders

Training may be either supervised or unsupervised, with the latter being the more common option. The activations in the top-layer may be considered as features and input into any appropriate classifier, such as a Random Forest, SVM, or other similar algorithm. The weights learned by training the layers separately are then utilised to initialise the weights in the final deep network, and the whole structure is fine-tuned beyond that point. The network may be fine-tuned using back propagation if an extra output layer is added on top of the network, on the other hand. Only if the parameters are started near to a suitable solution will back propagation operate effectively in deep networks, as previously stated. This is ensured by the layer-by-layer pre-training. There are also more ways, including as dropout and maxout that may be used to fine-tune deep networks.

### *1.8.3.2 Restricted Boltzmann Machine (RBM)*

There are two layers to RBM: the visible layer and the concealed one. Between layers, however, there are no links; only visible to concealed links are found here. It is trained to maximise the anticipated log probability of the data. For each input, a binary vector of Bernoulli distributions is used to learn the distributions. As with a conventional neural network, an activation function between 0 as well as 1 is often utilised, and the logistic function is used to calculate it. Assuming the output is random, all neurons are engaged if the activation is larger than the random variable for that particular neuron. Neurons in the buried layer get inputs from visible units. Initial input vectors for visible neurons are binary, and afterwards hidden layer probabilities are applied.

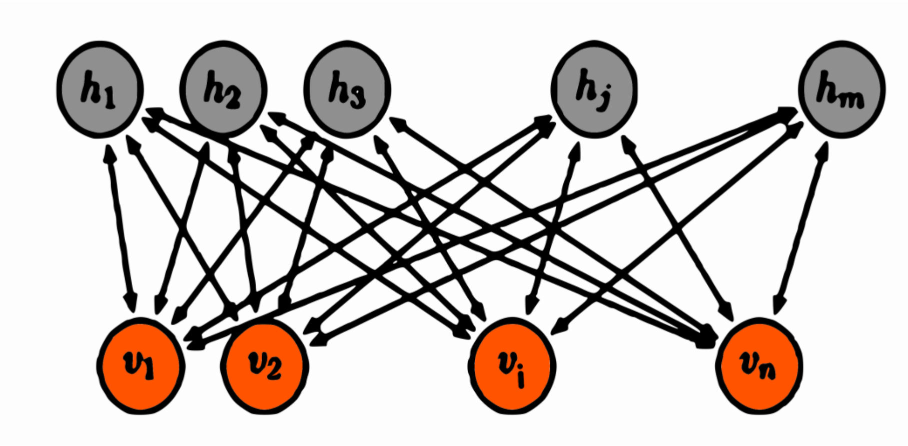


Figure 1.7: Visible and hidden layers in an RBM

#### *1.8.3.3 Convolutional Neural Networks (CNN)*

MLPs with a biological orientation are known as CNNs. In a typical Convolutional Neural Network, there are several layers of hierarchy, some of which are used to represent features and others of which are used as a form of traditional neural network for classification. Convolutional and sub sampling layers are two forms of modifying layers that may be used to change the appearance of images. There are two types of convolution layers: those that use equal-sized filter maps to execute convolution operations, and those that use smaller filter maps to accomplish sub sampling.

#### *1.8.3.4 Recurrent Neural Network*

Correctly put, an RNN model is a sequential model that is capable of portraying how successive elements relate to each other. Figure 1.8 depicts a feedback loop as a series of recurrences in both time and sequence. Exploding or disappearing gradients in RNNs are well-known issues. The learning process might come to a standstill if the gradient becomes too tiny or too great during the course of a lengthy data set. This problem was addressed by introducing the LSTM model and proposing the GRU model (gated recurrent unit). Both of these networks prevent gradients from bursting or disappearing by allowing the gradient to flow uninterruptedly across the network.

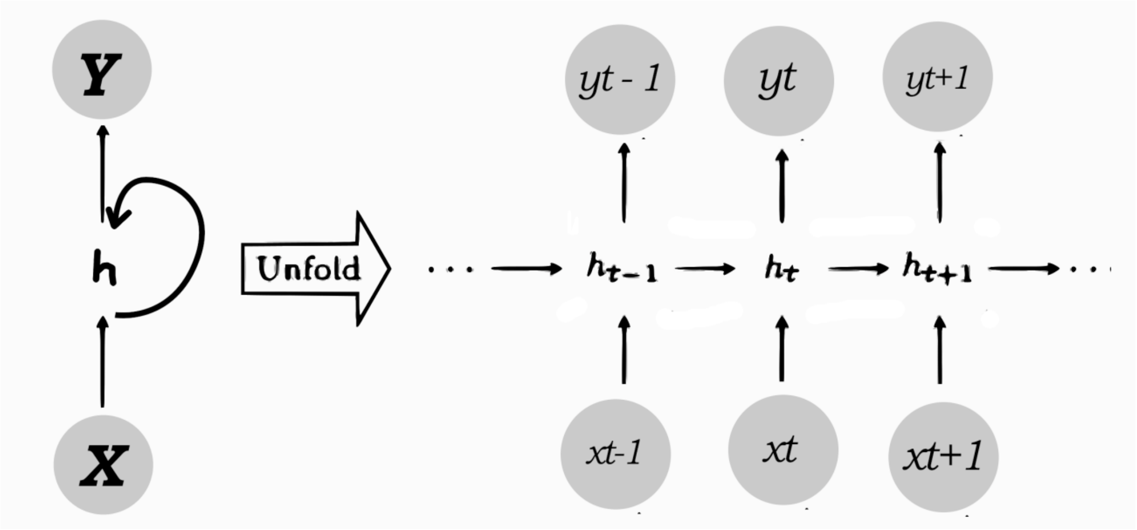


Figure 1.8: An illustration of a simple RNN and its unfolded structure through time t.

### *1.8.3.5 Deep Belief Networks*

DBN is a deep network architecture that uses Boltzmann-restricted cascading to generate the network layers. An RBM is trained to optimise the similarity between input and projection using a contrasted divergence approach (in terms of likelihood). DBNs are probabilistic frameworks because they prevent degenerate solutions by using probability. DBNs may be taught using a layer-by-layer greedy learning technique and then fine-tuned using gradient descent & back propagation methods such as SAEs. Employing a greedy approach of gradient descent and back-propagation, DBNs are constructed unsupervised and enhanced using layer-by-layer learning. (Vinyals et al., 2017) [37].

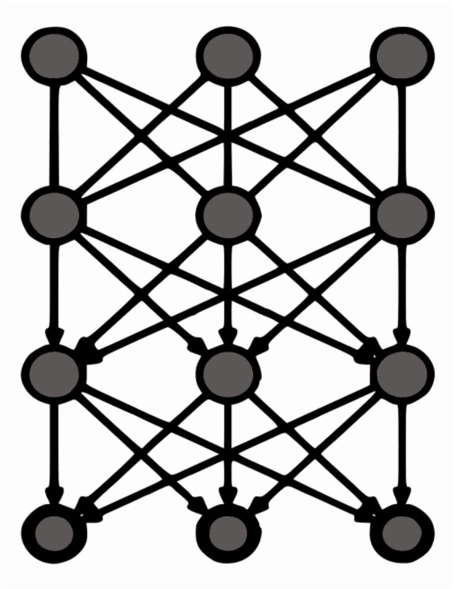


Figure 1.9: Structure of DBN

## 1.9 FEATURE EXTRACTION (FE)

Feature extraction (FE) from high-dimensional images is an example of dimensionality reduction, in which a large number of pixels from the picture are represented in a low-dimensional space so that the important sections of the image may be successfully recorded. Machine Learning, Image Processing, and more are all examples of its wide variety of uses. Pattern Recognition in image processing may make extensive use of this technology. Medical image analysis in particular, including counting red blood cells, white blood cells and cancer cell identification, has benefited greatly from the use of this technology. There are certain limitations in medical image analysis software feature extraction approaches due to the complexity of medical pictures. Colour (Gray) image characteristics, texture features, form features, as well as spatial linkages are often employed in feature extraction algorithms. The extracted image characteristics should be able to characterise things both abstractly and in a more precise manner. In order to do any computer vision, an image must exhibit properties including such distinctiveness, consistency; features are invariant under geometric shape, quickness, and abstraction. While several Feature Extraction techniques have emerged in the previous decade, each has its own pros and downsides(Kalaivani et al., 2020) [38]. Some are simple to implement, while others have low computational cost and quick processing speed, and so forth.

The method of removing valuable information from the raw data is called data mining. Nevertheless, for most FE approaches, the issue of extracting relevant features that may accurately capture underlying content of a piece of data or dataset as completely as possible continues to be a significant obstacle. It is the process of extracting the common characteristics from a picture so that it may be utilised for a range of different applications that is known as feature extraction. For the purpose of extracting visual characteristics, a variety of approaches have been developed. Image matching and recognition methodologies, as well as learning in machine learning techniques, can benefit from these characteristics(Khalid et al., 2014) [39].

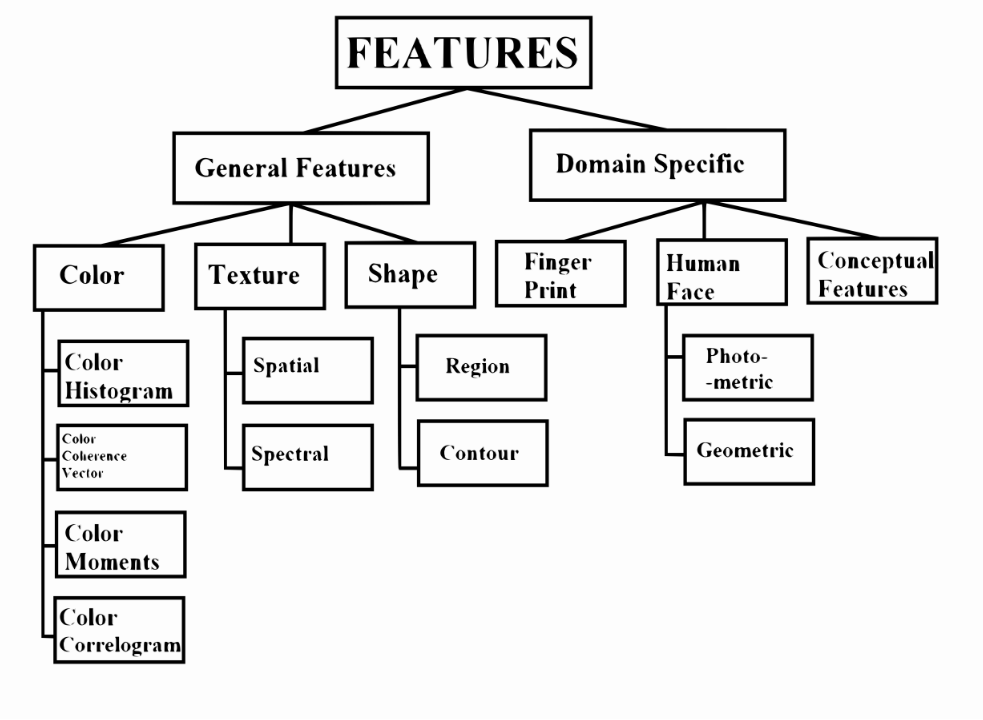


Figure 1.10: Classification of feature extraction techniques.

In order to execute image feature extraction on the component image, different picture pre-processing methods such as normalisation, thresholding, binarization, scaling, and so on are applied to the image. As seen before features may be divided into two categories: generic features (GF) and domain-specific features (DSF). Color, form, and texture are examples of application autonomous characteristics, while conceptual and human face aspects are examples of application dependant features.

### a) Color Features

In order to retrieve information that is presented in the form of a video or an image, colour features are often used for the extraction of visual characteristics. Color is most important characteristics of photographs, and it is characterised in terms of colour spaces or models, which are as follows: “RGB, HMMD, HSV, and LUV”. Color characteristics are resistant to changes in translation or viewing angle. This is to specify the various colour characteristics, and once the colour space for a given picture has been determined, the related colour features may be simply extracted from the image in question. It is possible to get several colour characteristics from published literature, such as colour correlation, colour histograms, colour coherence vectors (CCV), and colour moments (CM), to name a few examples. CM is the most straightforward and effective aspect of the group.

### b) Texture Features

Color features make use of individual pixels, while texture features make use of groups of pixels. Texture is used by the human visual system to aid in the interpretation as well as identification of images. Texture, in its most basic form, describes visual patterns that have the homogeneity quality. TF may be broadly classified into two types: spatial TF and spectral TF.

### c) Shape Features

When it comes to detecting and distinguishing real-world items, shape features (SF) play a critical role. They are the most dominating visual indication used by humans for similarity checking and matching. In general, SF are separated into two categories: region-based (RB) as well as contour-based (CB).

## 1.10 AIM OF THE PROJECT

- The aim of this project is to study the concept of clustering in depth with a twofold approach.
- The first step is concerned with the comparison of two well performing and popular algorithms namely K Means and BIRCH on five datasets.
- The better algorithm is applied for the application purpose. The better algorithm is used for image segmentation of CXR images in COVID 19 detection in the second step.

## 1.11 MOTIVATION

Clustering is very important concept in the upcoming fields of studies like ML, AI and other business applications. Primary motivation behind this study was to study clustering in depth with a twofold approach. In the first part the motivation behind the study is to deduce the better algorithm via testing it on five different datasets. Another motivation behind the application part is to implement the better algorithm deduced from first part & help the medical field, with using that algorithm for detection of COVID 19.

This research's motivation in the implementation part is to offer a consistent analysis system to support the decision-making of medical experts in the detection COVID-19 virus. Since the corona virus mainly attacks the respiratory system, the diagnosis of chest X-rays has emerged as a viable solution for detection of COVID-19 infection. RTPCR (Y. Fang et al., 2020) [40], on the other hand, is by far most successful method of COVID-19 detection. As a result of geological, sociological, and economic limitations, this procedure is very time consuming (taking hours to even days) and necessitates the use of specific kits that may not be accessible in distant sections of a nation. The fast antigen test, on the other hand, tests for the presence of viral antigens in a nasal swab, but it has a greater incidence of false negatives because of the higher likelihood of false positives. The serological test searches for antibodies created by the immune system against the virus in the patient's blood sample, which is collected during the procedure. In both CT scans and X-ray scans, invisible regions of the electro-magnetic spectrum are used in order to identify any kind of aberration. CT scans are employed for early detection and have a high level of clinical significance. In this article, we discovered that chest X-ray testing are both economically feasible and reasonably simple to interpret the findings of. Chest X-rays are readily accessible, are available in portable formats, and provide a minimal risk of radiation exposure. CT scans, on the other hand, are associated with a significant radiation risk, are costly, need clinical competence to perform, and are not portable. As a result, X-ray scans are more convenient than CT scans in many situations.

We have written two articles in which we will conduct a comparison research of the clustering algorithms and then the analysis and detection of the COVID disease through the K Means algorithm.

## 1.12 CONTRIBUTION OF THIS PROJECT

The following is the most significant contribution made by this paper:

- We have performed a thorough comparative evaluation on two well known and good performing clustering algorithms on five datasets.

- In the part 2 of our study, implement the K means algorithm for COVID-19 detection & Using our proposed framework, we tested the algorithms' validity on 3 well-known public X-ray and CT-scan image datasets, which were publicly available.
- Introduce an image segmentation model for segmenting COVID-19 CT images using a K-Means clustering approach, which demonstrated improved performance over previous models.
- The proposed Fast. AI framework was evaluated in comparison to earlier research on a variety of performance criteria, including precision, accuracy, recall, as well as f1-score. All measures have shown considerable improvements.
- After conducting a thorough assessment to verify the proposed approaches, we discovered that the proposed SVM classifier machine learning model has exceptional performance on 3-class classification tasks. After conducting a thorough assessment to verify the proposed approaches, we discovered that the SVM classifier model has good performance on similar studies. Studies have been conducted for the detection of brain tumour in MRI images, using SVMs, in which the linear kernel of SVM showed a very good accuracy of 91.66% [41].

Apart from medical imaging, SVM has also proved out to give good accuracy of 98.46% imaging in classification of plant diseases [42]. Brain tumour classification using images has also other studies where SVM has proven to be effective. Linear and Quadratic SVM has given 97.2% accuracy and cubic SVM has given 94.4% accuracy in previous studies [43]. Later on we discuss COVID 19 detection in previous studies using CXR images have also used SVM has given accuracies better than KNN classifier.

## 1.13 THESIS OUTLINE

This research work is partitioned into six chapters. The layout of every chapter is given beneath.

**Chapter 1** is the Introduction of this study. This chapter goes into detail on the first part i.e. the comparison of the clustering algorithms on 5 datasets & the second part on COVID detection.

**Chapter 2** is the study project's review of the literature. The purpose of this chapter to describe the studies and research results in the fields of clustering which relate to our first objective which is compare clustering algorithms & study ML, FE, & others in area of COVID-19 patient detection using CXR Images related works on the current study that have been conducted.

**Chapter 3** describes the tools used for implementation of this study.

**Chapter 4** describes the research methodology. The methodology of the two objectives which is comparison of algorithms and then implement and use clustering for analysis & detection of COVID are described in this section.

**Chapter 5** describes the outcome of the experiment. The proposed model of both the part of study is evaluated here and results are presented. Various evaluation metrics are used here.

**Chapter 6** is the concluding part of this paper. Conclusions of the studies and the potential future work are presented here.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 RELATED WORK

Clustering has become an important field of study for variety of studies in academics and Industrial applications. Numerous good studies have been conducted based on clustering and its applications in diverse fields. This section comprises of studying all the good studies related to our study which gives us inspiration to conduct our study and help us too.

We also study the good researches conducted related to the application of the clustering algorithm in our study which is analysis and detection of COVID 19. Since commencement of COVID-19 (Cov-19) pandemic, scientists have created ML-based algorithms to automated screening of positive COV-19 cases utilising a variety of radiological imaging, including CT scans and CXR. This section comprises research that have been conducted in relation to COV-19 diagnosis, and which have predominantly made use of artificial intelligence-based approaches, particularly ML as well as deep learning.

### 2.1.1 Based on Clustering

F.Wang, Hector-Hugo Franco-Penya, J. D. Kelleher, J. Pugh, R. Ross presented a study which contains a simplified version of Silhouette index, which was employed for validation of K Means method [14].

A study was conducted for Validation of the IRIS dataset by A. Vysala and Dr. J. Gomes. In this, the internal validation has been through the Silhouette Coefficient [44].

I. Ullah, H. Hussain, I. Ali, A. Liaquat have presented a study in which, K-Means algorithm was employed for Customer churn [9].

Y. Tu, Y. Liu, Z. Li have presented an article in 2010, in which the BIRCH Algorithm was employed with the focus on creating a new technique regarding time series segmentation [45].

An Enhanced model of K Means & Silhouette coefficient was previously employed in a study to detect the Dental Plaque has been presented in a study by P. Sudheera, V R Sajja, S. D. Kumar and N. G. Rao [46].

In an article presented by A.K. Singh, S. Mittal, P. Malhotra and Y. V. Srivastava, KMeans was applied to Cereal data, and to internally validate the result, “DaviesBouldin” Index was used [47].

A review study was conducted by A. Pugazhenthi and L.S. Kumar regarding the KMeans technique & Fuzzy C-Means [48].

The prediction of Customer churn in the telecom industry has been performed, and KMeans has been used in this study. This study was conducted by S. Preetha and R. Rayapeddi [49].

BIRCH technique has been applied in a study concerned with Malware identification, written by G. Pitolli, L. Aniello, G. Laurenza, L. Querzoni and R. Baldoni [11].

H. Mahi, N. Farhi, K. Labed and D. Benhamed have written an article in which KHarmonic Means with validation using silhouette coefficient was conducted with focus on clustering Multispectral Satellite images [50].

S.H. Jun and S.J. Lee, have conducted a study which compares algorithms using Internal Cluster Validation was also an inspiration for our article [51].

T. Gupta and S. P. Panda have conducted a study in which the IRIS dataset was taken up for the study, and used CLARA technique and K-Means technique for clustering. The internal validation was done by employing Dunn index and Silhouette Index [52].

A. D. Fontanini and J. Abreu conducted a study in which BIRCH technique was employed to extract the load profiles in a study [53].

M. Aryuni, E. D. Madyatmadja and E. Miranda have presented a study, in which, a customer segmentation study was performed, via K-Means and K-Medoids techniques [54].

A study was held in past by Y. Liu, Z. Li, H. Xiong, X. Gao and J. Wu, which helped us understand the validation measures. This helped us in understanding concepts and conducting our study [55].

K-Means Algorithm is employed in past study by X. Min and R. Lin, where fraud detection was done via signaling data [56].

The corpus callosum was segmented from brain magnetic resonance imaging using the K-Means approach by G.V. Bhalerao and N. Sampathila [57].

Medical Image segmentation has been performed through an optimized version of KMeans in an article by R. Alzu'bi, A. Anushya, E. Hamed, B.S. Angela Vincy and A. AlSha'ar [58].

In a previous study by S. Songma, W. Chimphlee, K. Maichalernnukul and P. Sanguansat, Two phase classification method was proposed, in which K-Means was taken up for clustering [59].

Previous study by T H Sardar et. Al, have proposed a document clustering method in which combination of MapReduce and K-Means have been used which have given better results [60].

In a previous study by Joy Iong Zong Chen, K-Means algorithm has been used with the CNN Algorithm for the recognition of vehicle license plates, which has also been improved by application of detection & location algorithms [61].

A comparative study between K-Means and Fuzzy C means has been performed before by K. Zhou et. Al which also aims at comparatively evaluating two different clustering algorithms [62].

Previously by C Cheng et al. a Fuzzy K-Means based control method has been proposed for improvement in boiler combustion efficiency of a 1000MW ultra supercritical power plant. GPC with combination of FKM has been used in this study to improve control efficiency [63].

K-Means has also been modified by repeated watershed transformation i.e. iterated watersheds which has shown improved results which has been illustrated through image segmentation by S Soor et. Al [64].

I Khan et. Al have previously modified Fuzzy K-Means by a twofold process, which could detect correct number of clusters while initializing different cluster centroids[65].

S De et. Al have demonstrated in a previous study that K-Means can be used in the segmentation of skin tones, which are under different illumination and backgrounds [66].

### 2.1.2 Based on COV-19 Diagnosis in CXR

A variety of studies have been published on COV-19 and how to identify it since emergence of COV-19 pandemic. These trainings include medical analysis, artificial intelligence, data mining, including data mining connected to COV-19. There are many CXR image datasets required to address COV-19 instances, & this literature review examines the existing CXR image datasets to detect COV-19 cases using different learning models that are applied to categorise the CXR pictures. Many evaluations have been conducted to highlight current advancements in recognition of COV-19 infection.

(Shorfuzzaman et al., 2021) [67] In this research, researchers present a unique CNN-based DL fusion framework that is built on transfer learning idea, wherever parameters (weights) after separate models are fused in single model to extract features from pictures, that are then fed to a bespoke classifier for predicting, utilising the transfer learning approach. The diseased regions in CXR images are shown applying gradient-weighted class activation mapping, which is a type of classification mapping. Also included is feature representation via visualisation, which permits us to obtain a improved grasp of class separability of the tested models in terms of COV-19 detection. When evaluating effectiveness of suggested models, cross-validation experiments are carried out using open-access datasets including healthy plus COV-19 & other pneumonia infected CXR pictures, respectively. It was discovered that best-performing fusion model can accomplish classification accuracy of 95.49 percent while also exhibiting a higher degree of sensitivity as well as specificity.

(Singh & Singh, 2021) [68] In this study, an automated technique for diagnosis of COV-19 using CXR pictures is suggested. Approach proposes an enhanced depthwise CNN for analysing CXR pictures, which is a breakthrough in the field. It is necessary to use decomposition method in order to incorporate multiresolution analysis into a network. Frequency sub-bands extracted from input photos are sent into the network, which then uses them to detect the illness state. The network is meant to predict if an input picture is normal, viral pneumonia, or COV-19 based on its classification. When paired with Grad-CAM visualisation, the forecasted output from the algorithm is used to provide a diagnostic. In addition, a comparison analysis with the current approaches is carried out. For the purpose of performance assessment, measures such as accuracy, sensitivity, as well as F1-measure are computed. The suggested technique outperforms the existing approaches in terms of effectiveness, and so as a outcome, it may be utilised to provide an accurate diagnosis of the condition.

(Krishnan & Krishnan, 2021) [69] The modern techniques of assessing chest X-rays as well as CT scans need extensive expertise and therefore are time-consuming, suggesting that they are spending valuable medical professionals' time at a period when people's lives are at danger, which is concerning. Using fine-tuning of the Vision Transformer, this research attempts to aid in classification of CXR, resulting in advanced efficiency in classification of CXR. CXR are applied in the suggested technique to develop pretrained models that are fine-tuned for identifying the existence of COV-19 illness. According to the accuracy score of 97.61 percent, precision score of 95.34 percent, recall score of 93.84 percent, as well as FL-score of 94.58 percent, this strategy is the most accurate method. This finding demonstrates the effectiveness of transformer-based models on CXR images.

(Ji et al., 2021) [70] This work presents a COV-19 detection approach that is based on the merging of picture modal features. Small-sample enhancement process, including such rotation, translation, as well as random transformation, are performed on chest X-rays in the first step of this approach. When it comes to extracting modal features, five traditional pretraining models are utilised. The model is trained as well as fine-tuned, and then it is evaluated using the ML evaluation standard, as well as the ROC is plotted against the model's performance. When compare to conventional model, the classification approach developed in this work is more successful in detecting COV-19 image modal data, so it getting the desired result of properly detecting occurrences.

(Peng et al., 2021) [71] Researchers performed multiple case studies to illustrate the usefulness of COV-19-CT-CXR. Researchers demonstrate that the inclusion of COV-19-CT-CXR as extra training data may result in enhanced DL efficiency for the categorization of COV-19 or non-COV-19 CT when combined with other training data. Secondly, researchers gathered CT pictures of influenza, another frequent respiratory syncytial ailment that may appear in a similar manner to COV-19. To establish the illness differences in research literature, researchers used text mining to extract captions as well as image explanations from research papers then contrasted 15 clinical symptoms as well as 20 clinical results of COV-19 to those of infection to display disease disparities. The dataset is one-of-a-kind in that it retrieves images together with pertinent text or fine-grained annotations, and it could be readily expanded in the coming. In addition to current sources, researchers expect that this effort will help to computer - aided diagnosis of the COV-19 epidemic.

(Hou & Gao, 2021) [72] this study, a novel diagnostic allows for the development on a DCNN has been created to help radiologists diagnose COV-19 pneumonia via separating it from non-COV-19 influenza in patients regardless of interpretation and detection of chest X-rays. The medical ability to identify and diagnose COV-19 will be improved by using an automated technology to speed up the interpretation of chest X-rays while also improving performance. X-ray dataset pictures are utilised in the DCNN to describe the behaviour of the training-learning models in order to get improved prediction performance by using the comprehensible technique. The average accuracy of technique is over 96 percent, that can substitute human reading but has possibility to be used to large-scale quick screening for COV-19 for extensively usage scenarios.

(Hastuti et al., 2021) [73] This research, the transfer learning model is used, allowing for an examination of the accuracy acquired and a comparison with the accuracy findings produced by the conventional learning approach. An open-source dataset of 1500 CXR images was utilised in this research. The precision of TL as well as selection of epoch or batch size parameters are used to gauge system efficiency. This shows that the optimal pair parameter utilising transfer learning techniques is paired by epochs 100, batch size 64, and yields a 98 percent accuracy rate. their transfer learning approaches outperform classical instructional strategies in terms of classification accuracy.

(Madaan et al., 2021) [74] In this study, researchers present a 2-phase X-ray image classification system, named XCOVNet, for early COV-19 identification utilising a cnn model, which is described in detail. It is possible to identify COV-19 poisons in X-ray pictures of patients in dual stages. Pre-processing 392 CXR pictures, half of whom are positive for COV-19 & another half negative. Patients may be classified with 98.44% efficiency after the second round of training and tuning neural network models.

(Calderon-Ramirez et al., 2021) [75] They use CXR pictures with uncertainty estimates to develop a COV-19 infection detection mechanism. Computer-aided diagnostic tools in the medical area must be used with caution if they are to be safe. A board-certified radiologist should review any model estimates that have a significant degree of uncertainty. Utilizing unlabeled data and the MixMatch semi-supervised approach, they hope to enhance uncertainty estimates. In this study, they evaluate the most commonly used uncertainty estimating methods, including Softmax scoring, Monte Carlo dropout, as well as probabilistic uncertainty quantifying. Jensen-Shannon distance here among uncertainty populations of right and wrong guesses may be used to verify the validity of uncertainty measurements. Just like many other prior measures, this one takes into account the distribution of uncertain estimates, unlike most other metrics. They found that incorporating unlabeled data improved their ability to predict uncertainties significantly. The Monte Carlo dropout approach yields the most accurate findings.

(Awasthi et al., 2021) [76] They are working on a mobile DL prototype for COV-19 diagnosis that is both lightweight as well as efficient, so this makes use of US lung pictures. COV-19, pneumonia, and healthy were the three groups that participated in this activity. The newly constructed network, dubbed Mini-COVIDNet, was contrasted to both existing lightweight as well as current heavy neural network models in order to determine its performance. The projected network has an 83.2 percent reliability rating and requires just 24 minutes of training. The Mini-COVIDNet that has been proposed uses 4.39 network parameters and just 51.29 MB of RAM, which is much less than the next most efficient network. A mobile application might be utilised to locate COV-19 care locations by using lung ultrasound imaging. When compared to other lightweight systems in general of accuracy & duration, the Mini-COVIDNet shown here outperforms them. This is in accordance with the distribution of numerous inconsequential systems on surrounded systems.

(Panetta et al., 2021) [77] The investigation is conducted out on double publicly obtainable datasets: (a) Kaggle & (b) COVIDGR, both of which are freely available online. For the purpose of evaluating the system's performance, a variety of evaluation metrics would be utilised. These measurements will include accuracy, a clear indication, precision, efficiency, as well as f1 scores. Kaggle x-rays revealed a difference of around

100 percent between normal as well as COV-19 x-rays when using the three-class categorization system described previously. When it comes to the specificities, they received an overall score of 77.72 8.06, whereas the COVIDGR dataset received an overall score of 72.65 6.83.

(Channa et al., 2020) [78] In this case of COV-19 pendamic, streaming diagnosis built on retrospective evaluation of laboratory data in form of CXR is required, and DL may be used to do this. A novel approach for detecting COV-19 was presented in this study, which included synthesising medical pictures with the use of deep networks. The research yielded good findings, with an accuracy of 91.67 percent in diagnosing COV-19 as well as an accuracy of I00 percent in determining the survival ratio.

(J. Liu et al., 2020) [79] this report presents COV-19 CT, a common image model for time-space sequences. This project successfully completed the creation of several deep models, ranging from CT segmentation in the pulmonary parenchyma of the lumbar spine to lesion identification that permits automatic case sensing of COV-19. By integrating images from various phases of a similar patient, it may be possible to acquire more accurate screening findings than previously possible. The COV-19 image detection framework for time-space sequences is suggested in this study as a standardising image detection model. The findings of the experiment demonstrated that the model is extremely accurate and reliable, with good support for diagnosis of COVID 19, and it has a favourable impact on epidemic detection and management in general.

(Haritha et al., 2020) [80] They make use of CheXNet algorithms to anticipate CXR data for this section of the new COV-19 disease, which is in its first year. With future improvement, this approach may be used in real-time COV-19 detecting circumstances, such as in traffic jams. Its key goal is to increase the number of data sets available in its framework in order to give better training for more exact estimations. During MLtraining, this improves the model's performance over a wide range of data sets, which is beneficial. It may be possible to predict the chances of this individual's survival in more detail.

(Ahsan Pritom et al., 2020) [81] The first comprehensive test was carried out in order to detect and identify websites that were infected with COV-19. A methodology was described, and it was then applied to a set of data. In their studies, they discovered that the attackers that carry out COV-19 hostile websites are agile, skilled, and driven by financial gain. If the WHOIS data for certain websites is available, their research has shown that an RF detector might be a useful tool in detecting this kind of assault.

(Kapetanović & Poljak, 2020) [82] The current coronavirus epidemic in the Republic of Croatia's Closed Territory has been patterned after a previous outbreak in the same region (COV-19). The revised SEIR prototype is designed to depict the pandemic by employing statistical numbers from the public, such as the number of people who were infected, recovered, as well as died during the outbreak. The SEIR model, which was implemented by the Croatian Ministry of Health at the beginning of COV-19 on February 25, provides introductory information about the future diseases, tighter monitoring of social distance, including quarantine action for those suffering with harmful illnesses. It will be possible to calculate the basic reproductive number if the information supplied is accurate and indeed the model used is accurate. The results hint to the possibility of significant improvements and urge on the Ministry of Health to take severe preventive measures.

(Yang et al., 2018) [83] In this study, a novel CT image denoisation approach built on Wasserstein distance and perceptual similarity is developed using a GAN. The Wasserstein distance is an essential notion in the best transport theory, and it has the potential to advance efficacy of GAN. Noise is reduced by associating the perceptual qualities of a designated output with perceptual characteristics of ground truth in a particular region, while GAN focuses on statistical migration from a powerful to a weak data distribution. As a result, suggested solution applies your visual perception experience to problem of picture degradation, and it is capable of not only dropping image noise, but of attempting to preserve vital information. They obtained promising results in their studies, which were carried out with the use of clinical CT pictures.

(Van Tulder & De Bruijne, 2016) [84] This work blends a generative and a discriminating learning aim, and it proposes a convolutional classification confined by the Boltzmann network as a solution. This aids in the learning of filters that are suited for the characterisation as well as classification of training data, which really is beneficial. They show CT pictures with lung texture and airway identification learning assessments, as well as other materials. When a mixture of learning objectives was used, the results were merely discriminatory or generative, with one example to illustrate an increase in performance of lung tissue categorization of between 1 percent and 8 percent. This demonstrates how discriminating learning may assist a learner who is generally unsupervised in the process of learning filters that are optimal for classification.

There have been various studies issued in the literature that have employed DL classifiers to recognize people infected with COV-19. The released datasets as well as DL models for CXR pictures for which many scientists have previously offered a survey have been deliberated in aspect.

(Dimas et al., 2021) [85] In this work, CNN architecture is used in conjunction with DL to recognise images. X-ray pictures are utilised to distinguish between people who are infected with COV-19 and those who are not. It is estimated that 2562 x-ray pictures have been taken, which have been split into two categories: positive as well as normal. It will also make use of CLAHE preprocessing, as well as two sets of data which will be utilised as DL training manuals, namely the original data as well as the CLAHE preprocessed data, for COV-19 x-ray picture. Training procedure is carried out with help of CNN and Resnet-101 framework. The research separated the data into two groups based on an 80:20 ratio of training data to test data. A comparison of classification performance using a confusion matrix reveals that the suggested technique has the greatest accuracy, 99.62 percent sensitivity, including 99.60 percent specificity, all of which are 99.61 percent.

(Rawat et al., 2021) [86] In order to address this issue, a COV-19 detection system has been developed. This method will aid medical specialists in interpreting the report, as well as the expert panel will then make a final decision. As previously noted, Deep Learning methods, particularly CNN, have shown to be great in medical picture interpretation & detection, & since our goal is comparable to medical image classification, CNN is an excellent candidate for our use case in this context. We investigated four alternative CNN architectures on a CXR for Cov-19 Analysis, which we found to be effective. The models that are being utilised have been pre-trained using ImageNet datasets. The findings have been obtained by the use of Transfer Learning. Using multiple architectures, a comparative assessment of the data revealed that structures based on CNN offer significant promise for Cov-19 diagnosis & recognition, according to findings.

(Nur-a-alam et al., 2021) [1] In this research, researchers used Convolutional Neural Networks to undertake thorough empirical analysis in order to identify such pneumonia on photos. Our examination of a collection of current CNN models reveals that some of these models are unsatisfactory decision-making capabilities. Cov-19 data from of the COVID-CXNet dataset as well as the Normal class data from the NIH CXR dataset, multiple binary class classification classifiers are based in this context. CHALE as well as BEASF based on the ASF were used to pre-process the data, as well as CHALE for the data. With the use of transfer learning approaches, the ImageNet pre trained modelling of different CNN models, such as DenseNet, VGGNet, and EfficientNet, are converted to this domain and used in other applications. By combining the EfficientNetB5 model with pre - processed data, the best possible result is attained, with an F1 score of 0.99.

(Akter et al., 2021) [87] Recent study has shown a correlation between presence of COV-19 and discoveries in CXR pictures. This paper's method builds on that research by using current DL models (VGG19 as well as U-Net) to analyse CXR images and categorise them as true or false for COV-19. Following the preprocessing stage, that also includes lung segmentation as well as removing the environmental factors that does not provide key data for the task and might even result in biassed outcomes, the developed scheme moves on to the classification algorithm trained using the transfer learning scheme, but instead eventually to the outcomes analysis and the findings stage, that also includes heat maps visualisation. The most accurate models had a COV-19 discovery accuracy of around 97 percent.

(Shadin et al., 2021) [88] The study offered two models for perceiving COV-19 from CXR images, one dependent on DL and another one on transfer learning, both of which were based on CNN. A total of 1553 CXR pictures were utilised in this study, which came from several datasets. Our suggested DL-based CNN architecture attained maximum training accuracy of 79.74 percent as well as the maximum validation accuracy of 84.92 percent. On the contrary, the TL-based InceptionV3 structure scored the greatest training accuracy of 85.41 percent and the maximum validation accuracy of 85.94 percent among all architectures tested.

(Wu et al., 2021) [89] This work describes the growth of a unique Joint Classification as well as Segmentation (JCS) algorithm for performing real-time as well as comprehensible COV-19 chestCT diagnosis in the laboratory setting. 144,167 chestCT scans of 400 COV-19 patients plus 350 uninfected cases were used to train the joint classification & segmentation (JCS) algorithm. COVID-CS dataset was used to train their JCS system. The suggested JCS diagnostic method for COV-19 feature extraction and classification has been shown to be very effective in extensive testing, according to the authors. This model achieves an average sensitivity of 95.0 percent and specificity of 94.0 percent on our COVID-CS classifications testing set, as well as a Dice score of 78.5 percent on the segmentation test set, using our COVID-CS dataset as input.

(Chang et al., 2021) [90] In this work, clinicians are assisted in identifying COV-19 illness from CXR pictures by using DL techniques Once the photos had been pre-processed, they were sent into the VGG16 model, which automatically classified them into three groups to aid radiologist in diagnosis & conduct of condition. The accuracy of categorization was found to be 78 percent, according to findings. Detail investigations revealed that by correcting imbalanced pictures issue, accuracy of system may be significantly improved. Furthermore, selecting most appropriate picture pre-processing methods has a great likelihood of producing superior outcomes.

(Morís et al., 2021) [91] Due to a lack of available pictures of this current illness, researchers describe in this paper novel ways for artificially boosting the dimensionality of portable CXR datasets for COV-19 diagnosis, which are in response to the poor supply of images of this recent disease. Thus, researchers merged three complimentary CycleGAN designs to do a simultaneous oversampling utilising an unsupervised technique as well as without the need for paired data, as opposed to the traditional approach. Although the portable X-ray pictures have low quality, we demonstrate inclusive accuracy of 92.50 percent in a COV-19 screening environment, demonstrating that they are suitable for COV-19 diagnostic tasks given the negative quality of the images.

(Lin et al., 2021) [92] Proposed AANet can adaptively retrieve typical radiographic observations of COV-19 from diseased locations with a variety of sizes & features are intended to address this issue. It is made up of two key elements: an adaptable deformable ResNet as well as an attention-based encoder, both of which are very effective. This network is intended to deal with the wide range of COV-19 radiographic characteristics. Numerous trials on a variety of publicly available datasets show that the suggested AANet improves current best practises.

(Z. Fang et al., 2021) [93] Effective control of COV-19 requires accurate identification at an early stage. Chest screening with radiographic imaging, furthermore to the RT-PCR swab test, plays an essential role in preventing the spread of virus. heir MSRCovXNet beats various government DL techniques in identification of COV-19 without the use of other DL methods. MSRCovXNet has an accuracy of 98.9 percent and recall of 94 percent without use of other DL techniques. When tested on the COVIDGR dataset, technique achieves an average accuracy of 82.2 percent, outperforming other methods by at least 1.2 percent on average.

(Arias-Londono et al., 2020) [94] This study gives an assessment of several deep neural network-based algorithms that were developed. In order to construct an automated COV-19 diagnostic tool that uses CXR pictures to identify among control, pneumonia, and COV-19 groups, the following first procedures must be taken. On the basis of a dataset consisting of over 79,500 X-Ray pictures obtained from various sources, the study discusses the procedure that was used to train a CNN. Three distinct experiments, each including 3 distinct pre-processing approaches, are carried out to assess and compare the models that have been built. The goal is to determine if preparing the data has an impact on the outcomes and whether it makes them more understandable. In the same way, a careful examination of various variability concerns that might jeopardise the system as well as its impacts is carried out. The applied approach achieves a classification accuracy of 91.5 percent, with an average recall of 87.4 percent for the poorest but most explainable test, which necessitates the use of an automated segmentation of the lung area.

(Sethi et al., 2020) [95] Discovered that a unique coronavirus spillover occurrence has evolved as a pandemic that is impacting public health throughout the world. The screening of a large number of people is necessary in order to slow the spread of illness in a given community. Real-time PCR a common diagnostic method for pathological testing. As a consequence, the rising frequency of misleading test results has created an opportunity for researchers to investigate alternate testing methods. CXR scans of COV-19 patients have shown to be valuable alternative signal in screening process for COV-19. Though, accuracy is dependent on radiological knowledge once again. Using a diagnosis recommender system to aid doctor in examining the lung pictures of the patients would lessen the diagnostic load placed on the physician. Medical imaging categorization has shown to be an effective use of DL methods, notably CNN. For diagnosis of COV-19, 4 distinct deep CNN architecture were tested on pictures of chest X-rays taken from various angles. It has been found that CNN-based architectures have the ability to diagnose COV-19 illness in certain cases.

(Bekhet et al., 2020) [96] This investigate describes an AI-based approach for early COV-19 detection using CXR pictures that makes use of medical knowledge as well as deep CNNs. to do this, a DL model is painstakingly constructed and fine-tuned to get best possible performance in COV-19 detection. Test data on current benchmark datasets reveal that the suggested method beats competition in terms of recognising COV-19 with a 96 percent accuracy rate.

(Jabber et al., 2020) [97] During the year 2020, a unique corona virus will have evolved as a highly pandemic illness that will damage public health throughout the globe. A significant number of individuals must be screened in order to determine those who are affected and prevent the spread of illness, which has become vital. There are failure scenarios for this tool because it produces more false testing results than necessary, necessitating the need to find a substitute tool. When it comes to COV-19 screening, CXR are a superior option to PCR. However, in this case, accuracy of findings is quite important. Researchers suggest a diagnostic recommendation system for reviewing lung pictures, which may aid physicians in their work and lessen the strain placed on them. An method based on DNNs. The CNN (convolution neural network) is utilised to achieve the highest level of accuracy possible.

(Liang et al., 2020) [98] An important goal is to develop a mapping between the visual patterns seen on normal chest X-rays and the COV-19 pneumonia CXR patterns. Although the original dataset has just 219 COV-19 positive pictures, it contains 1,341 photos of normal chest X-rays as well as 1,345 images of viral pneumonia, which is a usual unbalanced problem in the medical field. For the image-to-image translation, a U-Net-based architecture is used to produce synthetic COV-19 X-Ray chest pictures from regular CXR photos using the normal CXR imagery. To complete the final categorization task, an architecture consisting of 50 convolutional layers of residual net (ResNet) is used. When trained with the original unbalanced dataset, the model obtains an accuracy of 96.1 percent, which is higher than the accuracy of 95.6 percent achieved while trained using the train by-scratch approach. Furthermore, the classifier trained with the improved dataset exhibits more consistent measures of accuracy, recall, as well as F1 scores across multiple picture classes than the classifier trained with the original dataset. Finally, researchers accomplish that GAN-based data improvement technique is appropriate to the vast majority of medical picture pattern recognition tasks, and therefore it is a powerful tool for addressing the widespread problem of expertise dependency that exists in the medical field.

(Calderon-Ramirez et al., 2020) [99] This problem is addressed by developing a semi-supervised DL model that uses both classification and regression problems data. A semi-supervised DL system based on Mix Match architecture is designed and tested to categorise CXR into Cov-19, pneumonia, as well as healthy patients based on their appearance. The proposed technique was calibrated by 2 publicly accessible datasets, which were used in the calibration. The findings demonstrate that once labelled / unlabelled data ratio is low, accuracy increases by around 15 percent. This suggests that our semi-supervised approach may aid in the improvement of performance levels in the identification of Cov-19 once quantity of good labelled data is limited.

(Al Mamlook et al., 2020) [100] with this work, the researchers want to construct a model that will aid in classification of CXR medical pictures into normal (healthy) & abnormal (diseased) categories. As a result, to boost efficiency and accuracy, seven current cutting-edge ML approaches and well-known CNN models have been combined to obtain the desired results. In this work, we offer our DL for the classification problem that is trained with altered photos and goes through a number of pre-processing processes before being used. When comparing the DL approach for the classification task to the other 7 ML techniques, it was discovered that the DL technique performed the best. With only an overall accuracy of 98.46 percent in this research, researchers were able to accurately classify respiratory infection in CXR images using DL based on CNN. It had a more successful outcome in terms of identifying instances of Pneumonia.

(Tsai & Tao, 2019) [101] this study investigates the use of effective quantitative features for detecting pneumonia from Chest X-rays identifier by using CNN as well as decision fusion of feature selection for the detection of pneumonia from Chest X-rays identifier. For all 14 illnesses, we are able to identify those using Chest X-rays, and we get state-of-the-art findings for all 14 disorders. The average experimental findings obtain a high identification rate that is much higher than that achieved by the current known approaches on a consistent basis. Briefly stated, the suggested ML approach outperforms the strategies mentioned in the literatures as well as achieves a higher recognition accuracy rate of 80.90 percent for 144 layers of CNN, demonstrating potential applications in the classification of histology images from chest X-rays.

(Katona & Antal, 2019) [102] In this study, researchers offer a technique for analysing X-ray pictures that is lightweight, automated, and scalable. Using DCNNs, humans were able to detect the presence of 16 medical disorders after training the network. In our evaluation, researchers used publicly accessible data to demonstrate that their technique is similar to the advanced approach while using much fewer computing resources.

(Tang et al., 2019) [103] In this research, researchers offer an end-to-end infrastructure for abnormal CXR diagnosis that is built on generative adversarial one-class learning. In contrast to earlier techniques, our system only accepts standard CXR pictures as input data. As a result, it is possible to identify aberrant chest X-rays from normal ones. AUC of 0.841 is reached on the tough NIH CXR dataset in a one-class learning context, demonstrating the effectiveness of our technique. This has the potential to reduce the burden for radiologists, as shown by quantitative and qualitative testing.

(Kieu et al., 2018) [104] In this research, researchers present a DL model for detecting abnormal problems in CXR pictures using deep learning. The suggested model, referred to as Multi-CNNs, makes use of several CNN to decide on the input picture. Convolutional neural networks are used to construct each component of the Multi-CNN, which is built using the ConvnetJS package. The output of the suggested model is normal/abnormal density, as indicated by the name. This work also introduces Fusion rules, which are a way for combining outcomes of model's components and synthesising results of model as a whole. The experimental findings from our x-ray image dataset demonstrated the viability of a suggested Multi-CNNs model with a 96 percent success rate.

### 2.1.3 Based on ML Techniques

With the development of COV-19 illness around the globe, it has become a fascinating task for the whole human race to overcome. The accurate identification of persons infected with COV-19 is still a pressing global need, with an increasing number of cases reported each year. CXR images, among other things, are a promising tool for detecting COV-19 patients quickly and effectively. There have been several researches that have shown the effectiveness of employing ML classifiers in the analysis of COV-19 using chest X-rays. They carried out multiple contrasts among a subset of classifiers in order to determine which was the most effective.

(Jabra et al., 2021) [105] In this research, researchers study the possibility of using a mix of advanced classifiers in order to achieve the maximum precision and accuracy for detection of COV-19 from X-ray images. The researchers performed a detailed comparative analysis between 16 state-of-the art classifiers to achieve this goal. Innovation in this work comes from the methods we used to construct the inference system that enables us to identify COV-19 with high accuracy, which is detailed in the publication. Steps one, two, and three make up the methodology: (1) a comprehensive comparison of 16 advanced classifiers; (2) a comparison of number of items different classifiers, such as hard/soft majority, weighted voting, SVM, as well as RF; as well as (3) identification of optimal combination of DL models as well as ensemble classification method that results in highest categorization confidence on 3 classes. Research discovered that employing Majority Voting technique is an appropriate approach to use in over-all circumstances for this work, and that it can attain an average accuracy of up to 99.314 percent on a typical basis.

(Hasoon et al., 2021) [106] this work presents a technique for the categorization as well as early identification of COV-19 using X-ray pictures, which is accomplished by image processing. Several procedures, such as image pre-processing, feature extraction, as well as classification KNN as well as SVM are employed. A total of 5,000 photos are used in the testing of the six models, with a training percentage of 5 folds cross-validation applied to each sample. 89.2 percent to 98.66 percent accuracy in diagnosis was found in the assessment, indicating a high level of accuracy. This model exceeds other models in that it obtains an average accuracy of 98.66 percent, sensitivity of 97.76 percent, specificity of 100 percent, precision of 100 percent, and specificity of 100 percent.

(Haque et al., 2021) [107] The efficiency of SSL for COV-19 diagnosis from CXR pictures has been investigated in this study. ML approaches have been shown to be much more efficient and accurate than traditional methods for picture categorization. On either hand, supervised learning is often used methods in ML, and it is very handy for specialists in diagnosing and making informed judgments regarding COV-19. It is also one of the most effective approaches in ML. To perform supervised learning in classification task, a large no. of radiographic pictures with high precision are required, which may be a challenging problem in the medical area. A novel technique for COV-19 Diagnosis using a nominal dataset has been studied in order to solve the issue at hand. A prepossessing strategy, which involves collecting and integrating local phase image features into a multi-feature picture, has been examined for use in training our SSL model in the teacher/student paradigm. According to our findings, the SSL model achieves 93.45 percent accuracy when 17.0 percent of the complete dataset is used for training. In addition, we present comparative metrics of the SSL method in comparison to other fully supervised approaches.

(Chandra et al., 2021) [108] In this work, an automated COVID screening approach (ACOS) is proposed that apply radiomic texture descriptions from CXR images to distinguish between normal, suspected, & NCOV-19 people. Novel proposal proposes a two-phase classification strategy that leverages a majority vote-based classification - based composed of five benchmark supervised categorization approaches as the first phase and a majority vote-based classifier ensemble as the second phase. Following Friedman's post-hoc many contrasts and z-test statistics, it seems that the results of the ACoS system are statistically significant. In conclusion, results are compared to the most recent advanced methodologies that are presently available.

(Sheeba Rani et al., 2021) [109] This section presented a novel COV-19 diagnostic model that is built on variety of ML–based classification techniques on CXR.

The suggested technique gathers samples from patients' first using Internet of Things devices as well as transfers the data to a cloud server, where the real diagnostic is performed. The suggested model's performance has been verified using a CXR dataset to demonstrate its effectiveness. In study, it was discovered that AdaBoost with RF model outperformed competition with accurate results of 90.13 percent, F score of 90.28 percent, kappa value of 89.59 percent, as well as MCC of 87.44 percent compared to other models. When compared to similar approaches, obtained findings revealed that suggested model is more successful for the diagnosis of COV-19 in conjunction with streptococcus pneumonia.

(Izdihar et al., 2021) [110] This research sought to determine the sensitivity of CoV-19 detection using the CAD4COVID software, as well as the accuracy of classifier performance using Automated ML (Auto ML) method, to better understand virus. A total of 70 CXR pictures were evaluated, and 39, 20 and 11 patients were assigned a likelihood score in low range (0-35), medium range (36-65), and high range (66-100) categories, respectively. AutoML detection has a sensitivity of 0.99 and an accuracy of 0.83, according to results. In conclude, AutoML with best optimizer may be equivalent to CAD4COVID in terms of specificity and consistency in identification of Cov-19 in terms of precision or sensitivity.

(Silva & Fernando, 2021) [111] The study covers popular as well as publicly accessible labelled CXR datasets with their requirements and explores the labellers, labelling methodology utilised by them, as well as the labelling strategies employed by them in a complete manner. Next, common and successful image processing approaches for CXR pictures are discussed in further detail. The study then goes on to describe the various ML structures that are currently in use, as well as the usefulness of DNN for this good application. To wrap it all up, there is a discussion of the gaps and undiscovered areas that exist in the existing literature, as well as what the future may hold for them in ML-based automated pathology identification on chest X-rays.

(Meng et al., 2021) [112] This research describes a ML-based method for addressing the challenge of pneumonia diagnosis. Feature extraction as well as a dimensional reduction technique was used in the technique, which enabled it to improve classification performance while also lowering the problems associated with the training phase. The suggested methodology was tested against numerous mainstream deep neural networks in the assessment tests, and the results were compared with experimental results of the trials using our possible framework. Several deep neural networks for common terms, including as ResNet, MobileNet, as well as Xception, have been shown to surpass our suggested technique in experiments, according to the findings.

(Brunese et al., 2020) [113] In order to rapidly screen patients with the goal of detecting this new kind of pulmonary illness, they present in this study a technique that will automatically identify COV-19 disease by analysing medical pictures, which will be implemented in the near future. Researchers use supervised ML methods to advance a model built on a data set of 85 CXR that was made accessible for study purpose and is publicly available. The results of the experiment prove efficiency of the suggested strategy in distinguishing between COV-19 illness and other lung disorders.

(Yee & Raymond, 2020) [114] In this study, researchers present development of a pneumonia detection system based on feature extraction from CNNs. The feature extraction from CXR images was carried out with support of InceptionV3 CNN. Three classification algorithm models were trained using the extracted feature to predict instances of pneumonia from a Kaggle dataset, with the results being published in JAMA Internal Medicine. The three methods are KNN, NN, as well as SVM, and that they are described in detail below. Models were evaluated for their effectiveness using a confusion matrix that depicted the models' sensitivity, accuracy, precision, and specificity (sensitivity, accuracy, precision, and specificity). In the outcomes, it was found that the Neural Network model had the high sensitivity of 84.1 percent, following by SVM (83.5 percent) as well as the KNN (83.5 percent). The AUC for the Support vector machines model was 93.1 percent, which was the greatest of all of the classification techniques tested.

(Eljamassi & Maghari, 2020) [115] In this work, researchers current a classification model that can be applied to identify the presence of an infection in the CXR pictures taken. Patients with COV-19 were included in a dataset that included CXR pictures of healthy persons, those who had pneumonia due to SARS, streptococcus, or pneumococcus, or additional patients with COV-19. For the extraction of visual characteristics, the HOG is utilised. SVM, RF, and KNN are used to classify the pictures, with classification rates of 98.14 percent, 96.29 percent, and 88.89 percent, respectively, for the three classification methods. As a final recommendation, it is possible to detect COV-19 disease more effectively.

(Thepade & Jadhav, 2020) [116] This research seeks to develop an automated approach for identifying pictures of chest X-rays that have been infected with the Covid19 virus. The suggested technique makes use of a dataset that contains human CXR taken from non-infected persons & those who have contracted pneumonia or have been infected with the Covid19 virus. Again, for purpose of feature extraction, local binary patterns with changes in their input parameters are utilised. Measurements of findings reveal that ensemble of RTree-RForest-KNN provides best classification accuracy, & that ensemble approaches outperform majority of individual classifiers in terms of classification effectiveness. When evaluating input parameters of "LBP, parameters R=6 (P=48) and R=7 (P=56)" provide best speech for average of metrics for 10-fold cross validation in suggested Covid19 detection technique from CXR pictures.

(Thepade et al., 2020) [117] The current study presents a colour space-based global texture FE approach for identifying covid19-infected patients. Luminance Chroma characteristics of CXR pictures are retrieved from color spaces such as YCrCb, Kekre-LUV, & CIE-LUV, as well as from grayscale images. These extracted characteristics are utilized to train several ML classifiers & ensembles, which are then used to conduct 3-class classification as covid19, pneumonia, as well as normal, among others. The consequences of 10Kcrossvalidation reveal that ensembles outperform individual ML classifiers in terms of performance.

(Luo et al., 2020) [118] This research argues that adding an external CXR dataset results in poor training data, that increases difficulty of task. The faulty data is broken down into two categories: domain discrepancy, which occurs when the picture appearance differ across datasets, as well as label discrepancy, which occurs when various datasets are only partly labeled. In order to do this, we design the multi-label thoracic illness classification issue as a set of weighted independent binary tasks that are divided into categories. Considering common categories that are maintained across domains, researchers use task-specific adversarial training to mitigate the disparities in feature representations. They provide uncertainty-aware temporal ensembling of model predictions for categories that are available in a single dataset, which allows us to mine data from the missing labels even further. So that domain and label mismatches may be modelled and addressed concurrently, our system provides enhanced knowledge mining capabilities. Using three large datasets including more than 360,000 CXR pictures, researchers undertake a series of studies. We demonstrate that our technique beats other proposed model that establishes the advanced effectiveness on official NIH test set with an AUC of 0.8349, indicating efficiency of employing external dataset to enhance the internal classification usefulness of algorithm.

D. Narın and T. Ö. Onur (2021) In this research, suggested that a DL-based system that leverages CXR pictures from normal, COVID-19, and viral pneumonia patients be used to allow automated identification of COVID-19 patients to be used to identify the disease. Findings demonstrate whenever a DL-based model is employed in the original data, classification efficiency is 94.44 percent, as well as the maximum classification findings, is 82.30 percent when edge identification techniques are applied, respectively [119].

E. T. Hastuti et al. (2021) In this study, utilizing chest Xray images, the research presents the CNN approach for the identification of COVID-19 instances, which may be improved by using the TL model that can enhance accuracy even more. the Transfer Learning model DenseNet121 was employed. Results show that paired variables with epochs 100 and batch size 64 provide an accuracy rate of 98 percent, indicating that this is an optimal pair parameter when using TL approaches. [73].

S. Lafraxo and M. el Ansari (2020) In this study, researchers present CoviNet, a deep learning network that can be used to measure the levels of COVID19 in CXR images efficiently. Depending on AMF (Adaptive Median Filter), histogram equalization, as well as a convolutional neural network, the proposed scheme is a hybrid of the two.This model has an efficiency of 98.62 percent for binary classification as well as a precision of 95.77 percent for multi-class classification, according to results. Because early detection of COVID19 may help to prevent the virus from spreading, this paradigm can be utilized to aid radiologists in making the first identification of the virus [120].

E. Irmak (2020) In this study, a unique, strong, as well as resilient CNN model is constructed & suggested for the identification of COVID-19 illness utilizing publically accessible information. With an accuracy of 99.20 percent, this model may be used to determine if a particular CXR picture of a patient contains COVID-19 or not. The usefulness of the suggested approach is shown by experimental findings on clinical datasets [121].

D. F. Eljamassi & A. Y. Maghari (2020) In this work, researchers present a classification algorithm that can be used to identify the presence of an infection in the CXR pictures taken. For extraction of visual characteristics, HOG (Histogram of Oriented Gradients) is utilized. RF (Random Forests), SVM (Support Vector Machine), as well as Kclosest neighbors (KNN) are used to classify the pictures, with classification rates of 98.14 percent, 96.29 percent, and 88.89 percent, respectively, for the three classification methods. As a consequence of these findings, it is possible to identify COVID-19 illness more effectively [115].

In a 2022 study, A deep learning approach was taken to detect the COVID 19 cases via Chest X-ray images by A. Bhattacharyya, D. Bhaik, S. Kumar, P. Thakur, R. Sharma, and R. B. Pachori. In this study, a novel three-fold model was proposed for the aim, with segmentation using C-GAN, then extraction of discriminatory features using the deep neural Networks, and finally training the data on various Machine learning models. As a result, VCG-19 found the best results in this particular study [122].

A 2022 study by A. Hossein Barshooi and Abdollah Amrikhani used techniques by combining mainstream data augmentation techniques by Generative adversarial Networks and used different Filters. The Densenet-201 model was used to classify this data [123].

S. Sheykhivand et al In a 2021 Study, Multi-class classification was done using GANs with a combined technique of Deep transfer learning and LSTM networks. This study has also compared with other deep transfer learning models [124].

K. U. Ahamed et al In a study conducted in 2021 used CT Scans and Chest X rays images and a modified ResNet50V2 architecture to detect COVID-19 [125].

N. Rashid et al. In a study in the year 2021 was conducted to detect the COVID-19 Cases from the Chest X-ray images in a twofold CNN-based scheme. EfficientNet-B4 network applied in both the stages of this study [126].

## 2.2 SUMMARY

In general, DL and ML models for COV-19 identification that have been published in the literature have had trouble converging owing to the tough training approach they have through. The resulting model may not be stable or it may not contain the optimal set of parameters, as a result (weights). Therefore, the variety in training or validation losses will be greater than intended, as well as the loss will oscillate up and down throughout training phase as a consequence. As a response to this problem, we have developed a weighted fusion of variables inside the system on the basis of our solution. The weight training from numerous models of the communications system that were seen at the conclusion of the training process is taken into consideration by the investigators, who then take the average of those weights. Furthermore, while current ML methodologies have demonstrated superiority over other strategies in the detection of COV-19, the majority of these methodologies do not provide sufficient causal inference of models related to pertinent features of pathological signs in the CXR images to support their observations. It is thus impossible to determine the clinical effectiveness of these strategies unless more research is conducted to evaluate the high level properties retrieved from these models. Even if the experimental findings are quite precise, it is exceedingly unlikely that physicians in the actual world would accept a black-box categorization model. A technique for producing heat-maps that depict the regions of CXR pictures that are most suggestive of the illness is provided by our suggested system, which uses an explain-ability approach. Interpretability of our model's choice in a way that is clear to clinicians is provided as a result of this.

# CHAPTER 3

# TOOLS USED

## 3.1 PYTHON PROGRAMMING LANGUAGE

Python has achieved great appeal as an interpretive object-oriented programming language by developing complex software applications in several sectors like ML and DL. For the analysis of data, Python has a wide and robust collection of libraries and tools that are accessible for free. The term "big data" refers to the vast volumes of data that data scientists must deal with. Since it's simple to use and there are so many Python packages available, Python has become a popular choice for dealing with large volumes of data. ML, online services, data mining and classification, and other applications may benefit from Python's more robust analytics capabilities. Python, as a programming language, is quite forgiving and allows for pseudo-code-like programs. Useful for testing and enforcing pseudo code provided in instructional papers. Python has made this process rather simple in several cases. Python, on the other hand, is not perfect. The language is constantly created, and packages are notorious for their use of Duck writing. Using a package approach that yields something that looks like an array but isn't one may be annoying. Because the return type of a method is not specified in the standard Python documentation, this might lead to a great deal of trial and error testing that would not be necessary for a well-written language. Plus, this is an issue that makes learning a new Python package or library more complicated than it should be. Python as a programming language is supported by a large ecosystem. If you encounter any issues, all you have to do is go over to Stack Overflow and ask for help. Python is a widely-used programming language, which means that almost any question can be answered with a straightforward yes or no. Scientific computing is well-served by Python's extensive set of strong capabilities. NumPy, Pandas, as well as SciPy, are all free and well-documented packages. As a result of such packages, the amount of code necessary to create a particular program will be drastically reduced. This facilitates rapid iteration (Halterman, 2011) [127] (Jackson, 2006) [128] (Butwall et al., 2019) [129].

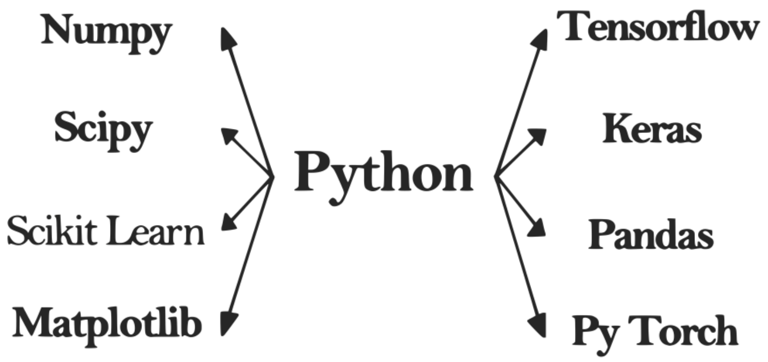


Figure 3.1: Machine learning in Python

### 3.1.1 Characteristics of Python

The following are some of the most significant properties of Python programming:

- Function and organized programming approaches, or object-oriented development, are supported.
- It may be used as a scripting language, as well as compressed to byte-code for the development of big programs.
- It offers identical high-level dynamic data structures & allows for dynamic type verification to be performed.
- It has the capability of supporting automated waste collection.
- It can be readily combined with C, C++, COM, ActiveX, CORBA, as well as Java, among other programming languages.

### 3.1.2 Applications of Python

As previously said, Python is extensively used programming languages on internet. I'm going to mention a couple of them below for consideration:

- **Easy-to-learn:** Python features a in significant no. of keywords, a straightforward organization, and a well-established syntax.
- **Easy-to-read:** Python code is better specified as well as apparent to the human eye than other programming languages.
- **Easy-to-maintain:** Python's programming language is quite simple to keep up to date.

- **A wide standard library:** The majority of Python's libraries are very portable as well as cross-platform interoperable with UNIX, Windows, as well as Macintosh operating systems.
- **Interactive Mode:** A feature of Python is its capability for the interactive environment, allowing enables for interactive development and testing of code snippets.
- **Portable:** Python is capable of running on a broad range of hardware configurations that have the same interface across all of these systems.
- **Extendable:** It is possible to include low-level modules in the Programming environment. These modules provide programmers the ability to enhance or adapt their tools to make them more effective.
- **Databases:** Each of the main commercial databases may be accessed with the Python programming language.
- **GUI Programming:** Python facilitates the development and porting of graphical user interface (GUI) programs to a variety of diverse calls, libraries, & monitoring systems.
- **Scalable:** Shell scripting gives better strength and structure for huge projects, while Python does not.

### 3.1.3 Python Libraries

There are several Python libraries available that provide strong and successful bases for assisting your data science work as well as the building of machine learning models. Whereas the list may well have been intimidating, there are a few libraries in particular on which you should concentrate your efforts since they are among the most widely utilized nowadays. Many libraries were used in this thesis, and they are listed in the next study framework (Beklemysheva, 2020) [130]:

#### *3.1.3.1 Numpy*

The major reason NumPy is utilized is because of its N-dimensional array capabilities. In comparison to Python lists, NumPy's arrays are 50 times more resilient. Some libraries, including TensorFlow, rely on NumPy for core tensor computations. Pre-compiled algorithms for mathematical procedures, which might be difficult to solve interactively, are also provided by NumPy. (Jan Erik Solem, 2012) [131].

#### *3.1.3.2 Pandas*

One of Python's most popular modules, Pandas is used largely for data analysis. Some of the most important tools for data exploration, cleaning, and analysis are included. All types of structured data may be loaded, prepared, manipulated, and analyzed using Pandas. Pandas Data-frames are used as an input by machine learning packages as well. (Santos, 2019) [132].

#### *3.1.3.3 Matplotlib*

Plotting tool Matplotlib is a Python library that allows computers to create a wide variety of graphing as well as visualization in a simple way. Jupyter Notebook's seamless integration with Matplotlib makes creating visualizations much easier than it was previously. In contrast to pandas, matplotlib does not use NumPy but rather python.

#### *3.1.3.4 OpenCV*

Since its inception in C++, OpenCV has been adapted to work in C++, Python, and now Java too though. It is a well-known object recognition package. The OpenCV website provides a free downloading of the software. This is a powerful tool for working with images, and it also has a broad spectrum of applications for image data editing, edge detection, and some other activities.

#### *3.1.3.5 Scikit-Learn*

Scikit-learn, a Python library for ML, is widely considered to be the most significant. To create machine learning models, you'll need to use scikit-learn, which has a plethora of tools for forecasting design and prediction, after having to clean or manipulate your data with Pandas or NumPy. Scikit-learn has a variety of uses. As an example, scikit-learn can be used to construct supervised as well as unsupervised ML models, cross-validated model accuracy, and perform feature importance analyses. (Van Der Walt et al., 2014) [133].

#### *3.1.3.6 Plotly*

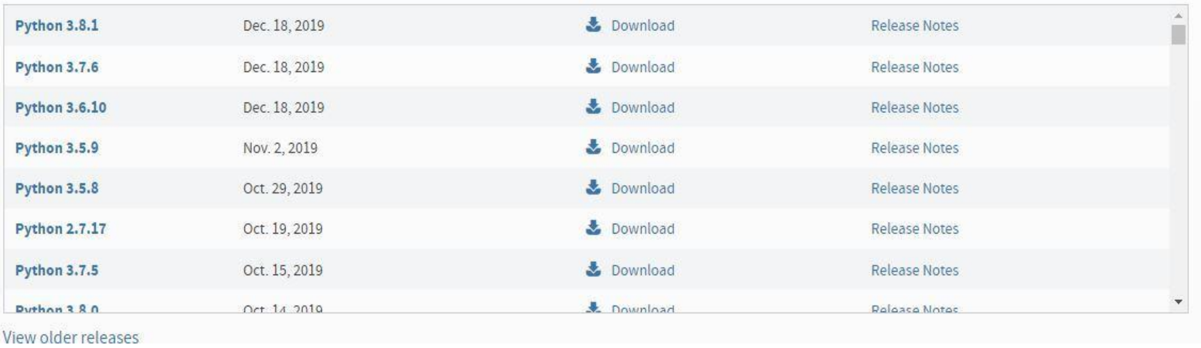
When creating visuals, Plotly is a must-have tool because of its high power, ease of use, and the ability to interact with the visualizations. Plotly and Dash, a program that uses Plotly visuals to create dynamic dashboards, go hand in hand. It is possible to execute these plots both online and in real life using Dash, a web-based Python interface that eliminates the requirement for JavaScript in analytical web-based applications.

#### *3.1.3.7 Seaborn*

Matplotlib-based Seaborn is a useful library for producing a variety of visualizations. Seaborn's ability to create data graphics that are enhanced is one of its most valuable characteristics. Data Scientists may get a better understanding of their models by seeing hidden relationships in a visual context. When it comes to data visualizations, it has adjustable themes and high-level interfaces that make it possible to create visually appealing charts that can then be exhibited to stakeholders.

### 3.1.4 Download and Install Python

You must first download the software before proceeding with the installation procedure. Python for Windows is accessible in all versions on python.org, which is where you may get them.



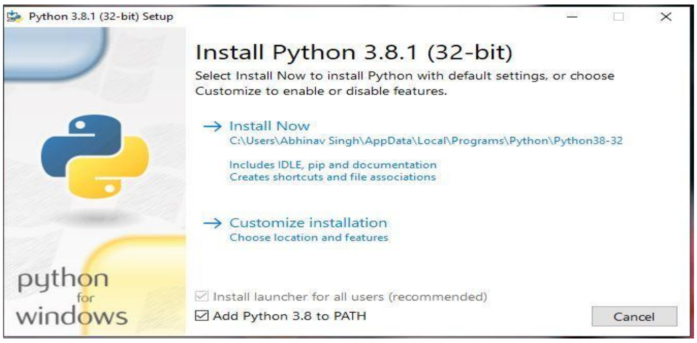
| | | | |
|---|---|---|---|
| Python 3.8.1 | Dec. 18, 2019 | Download | Release Notes |
| Python 3.7.6 | Dec. 18, 2019 | Download | Release Notes |
| Python 3.6.10 | Dec. 18, 2019 | Download | Release Notes |
| Python 3.5.9 | Nov. 2, 2019 | Download | Release Notes |
| Python 3.5.8 | Oct. 29, 2019 | Download | Release Notes |
| Python 2.7.17 | Oct. 19, 2019 | Download | Release Notes |
| Python 3.7.5 | Oct. 15, 2019 | Download | Release Notes |
| Python 3.8.0 | Oct. 14, 2019 | Download | Release Notes |

View older releases

Figure 3.2: Python Download

Download the necessary version as well as follow the on-screen directions to complete the installation procedure.

### A) Beginning the installation, Getting Started:



Install Python 3.8.1 (32-bit)

Select Install Now to install Python with default settings, or choose Customize to enable or disable features.

→ Install Now
C:\Users\Abhinav Singh\AppData\Local\Programs\Python\Python38-32

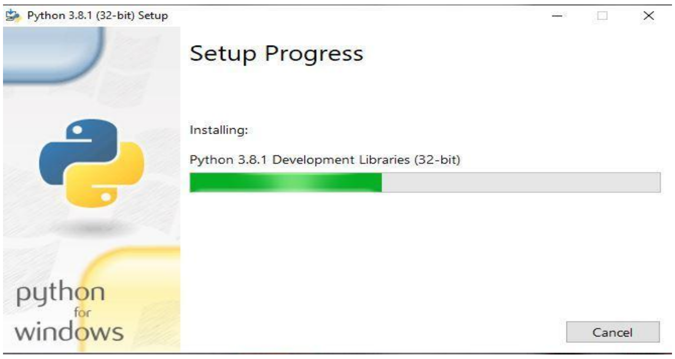
Includes IDLE, pip and documentation
Creates shortcuts and file associations

→ Customize installation
Choose location and features

☑ Install launcher for all users (recommended)
☑ Add Python 3.8 to PATH

Figure 3.3: Python Installation I

### B) Installing Libraries:



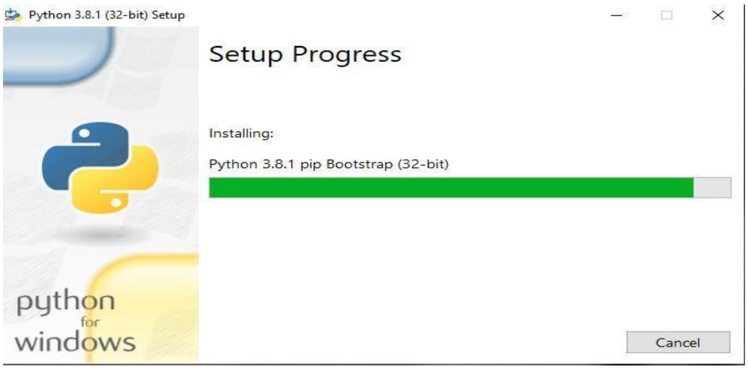
Setup Progress

Installing:

Python 3.8.1 Development Libraries (32-bit)

Figure 3.4: Python Installation II

### C) Installing pip and other features:



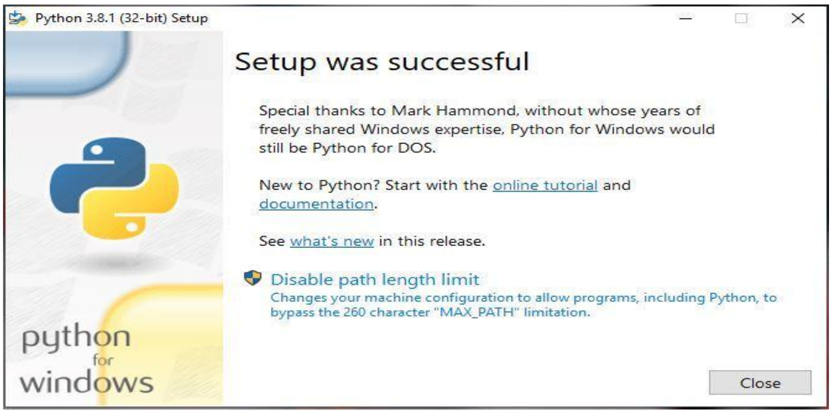
Setup Progress

Installing:

Python 3.8.1 pip Bootstrap (32-bit)

Figure 3.5: Python Installation III

### D) Finishing Installation:



Setup was successful

Special thanks to Mark Hammond, without whose years of freely shared Windows expertise, Python for Windows would still be Python for DOS.

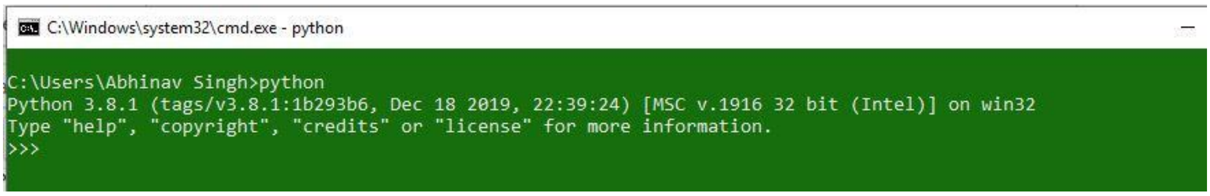
New to Python? Start with the online tutorial and documentation.

See what's new in this release.

Disable path length limit
Changes your machine configuration to allow programs, including Python, to bypass the 260 character "MAX_PATH" limitation.

Figure 3.6: Python Installation IV

Create the following command in your Terminal to check that the installation was successful.



```
C:\Users\Abhinav Singh>python
Python 3.8.1 (tags/v3.8.1:1b293b6, Dec 18 2019, 22:39:24) [MSC v.1916 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

Figure 3.7: Python Installation V

## 3.2 CODE EDITOR (JUPYTER)

Anaconda is the Python distribution that comes in second place in terms of popularity. If you want to install a third-party package, Anaconda provides an installer utility called conda that you may use to do this. Nevertheless, since Anaconda includes pre-installed using numerous scientific libraries, along with the Jupyter Notebook, you won't need to do something further than install Anaconda on your computer.

### 3.2.1 Jupyter Notebooks

The Jupyter Notebook is an open-source online tool that allows users to create as well as share information that includes live code, equations, visualization, and text. You may use it to generate representations of data code, equations, visualizations, and text. The Jupyter Notebook is developed by the members of the Project Jupyter community. Jupyter Notebooks are a spin-off program within the IPython project, which is utilized to have its own IPython Notebook project. Because it supports three programming languages in particular: Julia, Python, as well as R, the name "Jupyter" is derived from these three languages. Jupyter comes pre-installed with the IPython kernel, which enables you to write your programs in the Python programming language. However, there are presently over 100 different kernels available for you to choose from.

The Jupyter Notebook is an open-source online application that enables you to create custom notebooks which feature live code, equations, graphics, or descriptive prose. It is free to use it and open-source software. It is completely free though and can be obtained by visiting the Jupyter webpage. This includes data purification as well as transformation, simulation methods, multidimensional statistical study, data visualisation and ML, as well as a wide range of additional functions.

Three components make up the Jupyter Notebook in its most basic form (Northwestern, 2019) [134]:

- It is an **interactive web application** that enables you to write as well as run code in real-time while concurrently writing notebook papers. It is free to use.
- **Kernels**: In the notebook's web-based application, the kernels are distinct operations that are launched by the notebooks web service that run users' code in a particular language. The response is delivered to the notebook's web-based application where it was first entered. Also included are duties such as computations for interaction gadgets, tab completion, as well as transparent management in the kernel.
- **Notebook documents**: A self-contained report is a document that contains recognition of all substance demonstrated in the notebook's web application, including connections of calculations, narrative prose, equations, photographs, and interactive rich media visualizations of elements. These manuscripts are referred to as single docs. There is a kernel for every notebook document that is created. Users have the option of exporting the notebook in several various formats, like Latex and PDF, among others.

# CHAPTER 4

# RESEARCH METHODOLOGY

## 4.1 PROBLEM STATEMENT

Our aim is to study concepts of clustering in depth, and in a twofold approach with first performing the comparative study and then the application study. Since clustering has become such an important field and is being used in a wide area of academic research and industrial applications, we intend to compare the K Means Algorithm and BIRCH algorithm using 5 Datasets. The results derived from this study would be used in the application study, where we will use the better clustering algorithm for the image segmentation for CXR images and detect COVID 19. COVID-19 outbreak, which has touched approximately 216 nations & territories throughout the world, has thrown the world into chaos. There has been an increase in interest in computational model-based diagnostic techniques to assist with COVID-19 case screening and diagnosis via medical imaging, like CXR scans, since the outbreak of the disease. The CXR images of patients who have been infected with COVID-19 contain aberrations that show different radiological patterns, according to early findings. Even for experienced radiologists, detecting these patterns is difficult & time-consuming. With the considerably bigger dataset, we noted that the model appeared to have an over fitting problem and the performance somewhat deteriorated). However, the practical use of these strategies is unknown until more research into the high-level characteristics collected from these models is conducted. Although some studies imply that CXR scans aren't the greatest tool for detecting COVID-19 early, other studies (published in the literature) demonstrate that radiography images provide important information regarding COVID-19 infection as the disease progresses.

## 4.2 PROPOSED METHODOLOGY FOR COMPARATIVE STUDY

Here, methodology of the study is described in three sub sections.

### 4.2.1 Dataset Description

In this section the detailed description of the datasets is given on which the two clustering algorithms are compared. In this part, total of five datasets were used. Four datasets are collected via the website Kaggle, and one is available in the ‘seaborn’ library of the Python programming language. Table 4.1 depicts brief description of all five datasets.

Table 4.1: Brief description of Datasets

| DATASET | ROWS | COLUMNS | SOURCE |
|---|---|---|---|
| Dataset 1 (D1) | 1002 | 23 | Kaggle |
| Dataset 2 (D2) | 848 | 5 | Seaborn Library of Python |
| Dataset 3 (D3) | 200 | 5 | Kaggle |
| Dataset 4 (D4) | 2240 | 29 | Kaggle |
| Dataset 5 (D5) | 420 | 12 | Kaggle |

D1 is regarding the Software Architecture styles [135]. This data, as already indicated, is of shape 1002 x 23. D1 was secured from the Kaggle website. D1 contains attributes like Timestamp, Name, Organization, Last Degree, Job Experience , Repository Architectural Styles used, Client Server Styles used, Abstract Machine Styles used, Object Oriented Styles, Function Oriented Styles used, Driven Styles used ,Layered Styles used, Pipes & Filters Architectural Styles used, Data Centric Architectural Styles used, Blackboard Architectural Styles used, Rule Based Architectural Styles used, Publish Subscribe Architectural Styles used, Asynchronous Messaging Architectural Styles used, Plug-ins Architectural Styles used, Micro-kernel Architectural Styles used, Peer-to-Peer Architectural Styles used, Domain Driven Architectural Styles used, Shared Nothing Architectural Styles used. One hot encoding was performed on two of its attributes. D2 is secured from the Seaborn library of the Python programming language. This Dataset 2 has attributes like align, choice, time, coherence, firing rate. One hot Encoding was performed on two of its attributes.D3 was collected from Kaggle, and is named Mall customer Segmentation Data [136]. This dataset contains attributes of ID of the Customer, Age, Gender, Income Annually and Spending Score. Label Encoding was performed on the attribute of Gender in this dataset.D4 was collected from Kaggle, and is named Marketing Campaign Dataset [137]. One hot encoding on two of its attributes was performed.D5 is also collected from Kaggle; the “Birds Bones and Living Habits” Dataset, and has attributes like, Length and the Diameter of the “Tarsometatarsus”, Length and Diameter of the “Tibiotarsus” , Length and Diameter of the “Femur”, Length and Diameter of the “Ulna”, Length and Diameter of the “Humerus” [138].

### 4.2.2 Workflow of the Study

In Figure 4.1 explains the workflow of our study conducted in our article. After data cleaning and pre-processing, principle component analysis is performed. After that, both the clustering techniques are performed. Finally, Comparison is made.

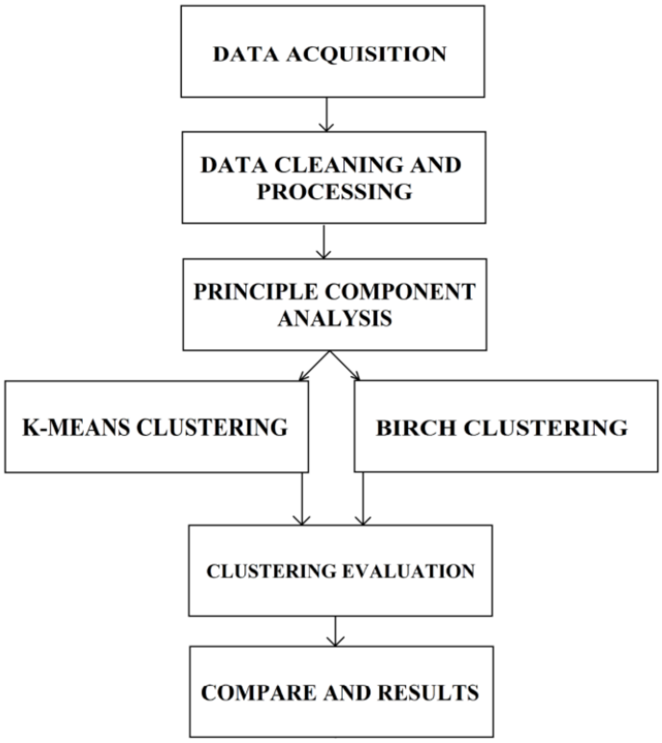


Figure 4.1: Block Diagram of Clustering study

### 4.2.3 Validation Metrics, K Means & BIRCH

Here, clustering is performed via K Means cluster technique and BIRCH clustering method on the dimensionally reduced dataset, which has two newly extracted features.

K Means is a type of clustering algorithm that uses an objective portioning criterion to construct clusters. K Means is a clustering method established on centroid values.

- The centroid of a cluster is used to illustrate the cluster in a centroid-based partitioning technique.

- The centroid of a cluster is determined as the average of the data points that make up the cluster using the K Means approach.
- Then it chooses 'x' of the objects in the dataset at random.
- For the remaining items, the Euclidean distance among the object and the cluster mean is used to assign the object to the cluster to which it is most comparable [27]. This improves the with-in cluster variation by computing new means using objects allocated to a cluster in last iteration. The algorithm has been taken from Data Mining: Concepts and techniques by J Han et. Al [27] where full algorithm has been given. BIRCH stands for Hierarchical Method of Clustering, and it operates via assembling data objects into a tree of clusters or a hierarchy. BIRCH is a two-phase process.
- The Clustering Feature Tree (C.F. Tree), is constructed in the first phase.
- In the second phase, BIRCH clusters the leaf nodes of previously developed tree using a clustering algorithm, which aggregates dense clusters into larger ones and removes sparse clusters as outliers.

The techniques carried out in this particular article are validated by internal validation metrics. Internal validation metrics are basically used to perceive the quality of

results obtained after clustering. The algorithm has been taken from [27].

## 4.3 PROPOSED METHODOLOGY FOR APPLICATION OF K-MEANS ALGORITHM

To solve the aforementioned difficulty, we have improved ML approaches for high-accuracy identification of COVID-19 utilizing CXR pictures. We gathered CXR pictures from 2 publicly available open-source repositories. The author's GitHub repository has a total of 616 positive COVID-19 pictures, all of which were acquired from author's repository. Next, we got the similar no. of normal and other non-COVID19 pneumonia images from the Kaggle collection to compare outcomes of 2 groups. Our research included exploring and analyzing data sets, pre-processing approaches, feature extraction, segmentation methods, and classification methods, among others. Firstly, apply the pre-processing techniques using Image filtering using an unsharp method and Image resizing into 196 * 196 after this apply the segmentation method that is K-means clustering and then use feature extraction which is extracting the feature with the help of principal component analysis, lastly apply classification technique. Then, the PCA is combined with the consolidated machine learning method which is the SVM classifier. The suggested approaches provide better outcomes for COVID-19 diagnosis than other methods that have been published before. Radiologists & virologists, in our view, will benefit from our efforts in fight against the COVID-19 epidemic since they will be able to make better and quicker diagnoses.

### 4.3.1 Data Pre-Processing

When it comes to Image pre-processing, it's described as the process of converting or encoding data in such a way that it can be readily interpreted by computers to provide reliable information. To put it another way, information should be converted in such a way that it can be readily understood by multiple algorithms, resulting in more accurate outcomes in the process. It is not required for every dataset to have full and accurate pure data. To get relevant information and reliable classification, image processing must be performed on each picture to remove noisy or distorted pixels. Remove noisy or distorted pixels from each picture to get more information and more accurate classification. After that, the images were transformed from RGB to greyscale using python program, and then they were resized to 224x224 pixels so that they could be loaded into the database for use by the computer system. Resizing means changing the size of the image, whatever the method: can be cropping, can be scaling.

#### A) Image Filtering Using Unsharp Method

Filtering methods save important information in a picture while filtering out any noise that may be there. This procedure may be viewed by taking a look at the many filters that are used in the process (Zaafouri et al., 2011) [139]. The unsharp filter is a simple and direct sharpening operator that derives its name from the fact that it improves edge sharpness (as well as other high-frequency components in an image) by performing a process that deducts an unsharp (or smoothed) version of the image from the original image, resulting in sharper edges.

#### B) Image Resizing

Since most photos aren't the actual size we need, it's essential to understand how to resize an image correctly. The pixels in an image are updated when it is resized.

Resizing means changing the size of the image, whatever the method: can be cropping, can be scaling.

### 4.3.2 Image Segmentation (IS)

IS (Y. Song & Yan, 2018) [140] is a technique of dividing an image into numerous no. of segments, and it is a way to accurately identify pixels of an image in a decision-oriented application. In the context of IS, this procedure refers to the process of splitting a picture's elements into desirable segments or portions that share features such as texture, intensity, and pixel value. So we can conclude (Ghoniemy, 2016) [141] that the objective of IS is to shorten or modify the representation of a picture, as well as to transfer the information contained inside an image into a more understandable form, to make the future analysis simpler. COVID-19 detection is more effective when it is split into smaller segments. To attain high segmentation effectiveness, a clustering approach called K-means clustering was developed to be used.

#### A) K-means clustering

It is an unsupervised image segmentation approach to use the clustering method (Kim et al., 2020) [142]. The K-means clustering method (H. M. Liu & Lu, 2015) [143] is a widely used clustering technique. Within a data vector, the K-means algorithm divides the data vector into predetermined no. of groups. Centroids of predetermined clusters are initialized randomly at the start of the simulation. When k-means clustering is used, a data collection is separated into a set of data with a k no. of groups, which is then divided again. Each cluster in the partition is distinguished by the data member it contains as well as the cluster centroid it contains. As a general rule, the centroid of a cluster may be defined as the point at which the total distances between all elements in the cluster are the shortest. The K-means (Eamani et al., 2020) [144] algorithm, as a consequence, is an iterative strategy that is aimed to minimize the sum of distances between each object, as well as the centroid of every cluster, throughout all clusters.

### 4.3.3 Feature Extraction (FE)

FE is a concept that is often used in computer vision and IP applications. FE (F. Song et al., 2010) [145] is a kind of dimensionality reduction that provides more information about the source picture than other forms of reduction. To be processed, input data must first be translated into a smaller representation set of features. FE is the term used to describe this process. Features are extracted from images to divide them into distinct subsets, which are often solitary points, a continuous curve, or an area (Kaur & Sharma, 2019) [146]. PCA is often used for FE in CT-image datasets to identify COVID19, and it is particularly effective for this purpose.

#### A) Principal Component Analysis

PCA (Kristiyanti & Wahyudi, 2017) [147] is a well-known and very effective approach that is used in image identification and compression for the extraction of features as well as the representation of data. In area of pattern recognition, computer vision, and signal processing, this approach is frequently employed. The main goal of PCA is to decrease the high dimensionality of the input space (spectra were recorded) to the significantly poorer dimensionality of the input space (independent variables), that is necessary in order to accurately portray the information. The presence of a statistically significant correlation between observed variables shows that this is true. PCA successfully decreases the number of features in the data set by removing minor components & shows the data set in a low-dimensional subspace (Suganthy & Ramamoorthy, 2012) [148].

### 4.3.4 Classification

Classification is an ML method that is used to predict which data examples belong to which groups. Machine learning is the term used to describe a system that has the potential of autonomously learning knowledge from experience and other sources of information.

Generally speaking, a classification algorithm is a function that balances input features in such a way that the output separates them into 2 classes: one class with positive values and also another class with negative values (or vice versa). It is developed by comparing the sensitivity and specificity when the threshold of the distance from the classifier boundary is changed over time (Erdaw & Tachbele, 2021) [149]. COVID-19 patient detection using an SVM classifier for the classification. It is most often used for classification in SVM. According to this approach, each data item is plotted as a point in an n-dimensional space, with each characteristic's value reflecting the value of a certain coordinate.

### A) SVM Classifier

The SVM classification method is explored in detail in this section. SVMs are a subset of the wider category of kernel algorithms. Using dot-products, a kernel method is an algorithm that is solely dependent on the data it is given. When this is the case, the dot product can be substituted with a kernel function that calculates the dot product in some feature space, which may be of high dimension. When it comes to identifying and categorizing flaws, SVM is a well-known approach. Because of its superior performance and ability to execute experiments well, SVM is becoming more popular.

To effectively translate the training data into a higher dimensional vector space, SVM makes use of non-linear functional mappings. As a result, the technique seeks for an ideal hyperplane to separate classes in question by ensuring that the observations obtained near borders of hyperplane, called support vectors, are segregated in the most optimal way possible and therefore maximize the marginal distance. Even though these support vectors are difficult to categorize, they, when combined with the outliers, give critical information concerning classification (Chen et al., 2018)[150](Mago et al., 2016)[151]. To put it simply, the essential concept underlying SVM is the development of an ideal hyperplane that can then be utilized for classification of linearly separable patterns. In pattern classification, an ideal hyperplane is a hyperplane chosen from a collection of hyperplanes for categorizing patterns in such a way that it maximizes the margin of the hyperplane, which is the distance among hyperplane and the closest point of each pattern. The most important goal of SVM is to maximize the margin for it to properly categorize the provided patterns; that is, the higher the margin size, the more accurately and the patterns are classified by SVM.

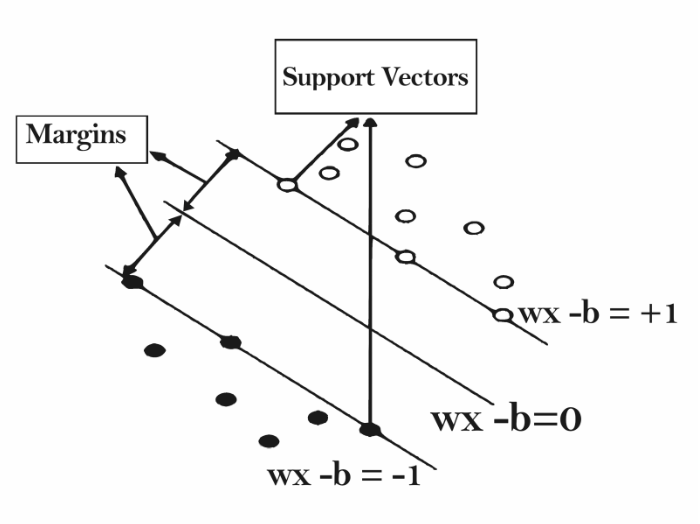


Figure 4.2: SVM Model

The support vector machine approach is shown in Figure 4.2, which is a straight forward model. The model is comprised of two distinct patterns, and the purpose of SVM is to distinguish between the two patterns in the model. The model is made up of three main lines of work. The line “w.x-b=0” is referred to as the margin of separation or marginal line in mathematics.

## 4.4 PROPOSED WORKFLOW

**Step 1.** Start

**Step 2.** Gather Dataset (CXR images).

**Step 3.** Input COVID-19 positive images.

**Step 4.** Apply image processing using Image Filtering(Unsharp Method) and Image Segmentation (K-means).

**Step 5.** Feature Extraction using PCA.

**Step 6.** Split dataset into training and testing.

**Step 7.** Train the model with SVM classification.

**Step 8.** Measure performance.

**Step 9.** A compared model with existing approaches.

**Step 10.** End

## 4.5 PROPOSED FLOWCHART

Here, in this section, we provide a proposed flowchart in Figure 4.3 for this implementation work. Each step is described in the proposed methodology section as well as the result section. once it gets trained then we have followed different steps to sentiment analysis as mentioned below diagram:

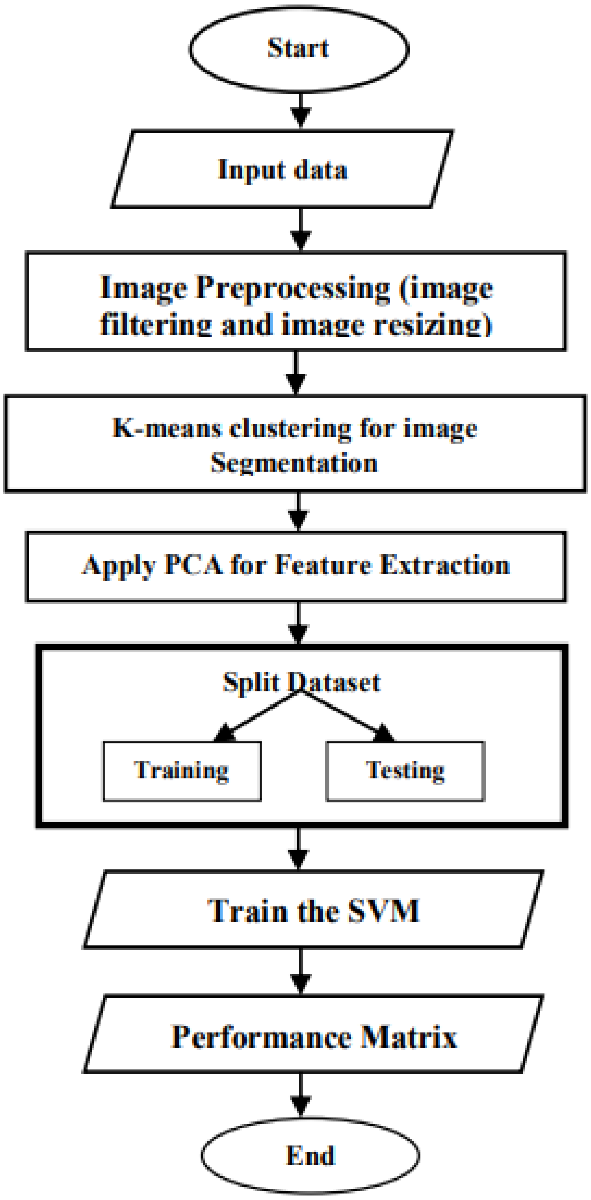

Start
Input data
Image Preprocessing (image filtering and image resizing)
K-means clustering for image Segmentation
Apply PCA for Feature Extraction
Split Dataset
Training
Testing
Train the SVM
Performance Matrix
End

Figure 4.3: Flow Chart of proposed Methodology

The suggested work is shown in Figure 4.3. The first stage is to pick the dataset for evaluation and comparison; the second step is to do pre-processing on the dataset. The third phase is picture segmentation, which is accomplished via the use of K-means clustering. The fourth phase is feature extraction, which is accomplished via the use of the PCA approach. Then there's a data split that leads to two-part training and testing. The third and fifth algorithms are classification algorithms, which are used to compare and contrast various methods. In the next stage, performance is evaluated and algorithms are compared to determine which algorithm is the most accurate in terms of accuracy.

# CHAPTER 5

# RESULT ANALYSIS

## 5.1 RESULTS: COMPARATIVE ANALYSIS OF PERFORMANCE OF ALGORITHMS

We present the results of the first part in this section. The result is explained in three brief sections

### 5.1.1 Dimensionality Reduction using PCA

Here, application of Principal component Analysis (P.C.A.) for dimensionality reduction is done and new features are extracted. By projecting the data into a much smaller space, PCA decreases the data's dimensionality. Rather of selecting features, this method creates a union of attributes by establishing a smaller collection of variables. Figure 5.1 shows the visualization of five datasets. Each graph in the Figure depicts the five datasets after dimensionality reduction. First feature is represented on the x-axis and second feature is represented on the y-axis.

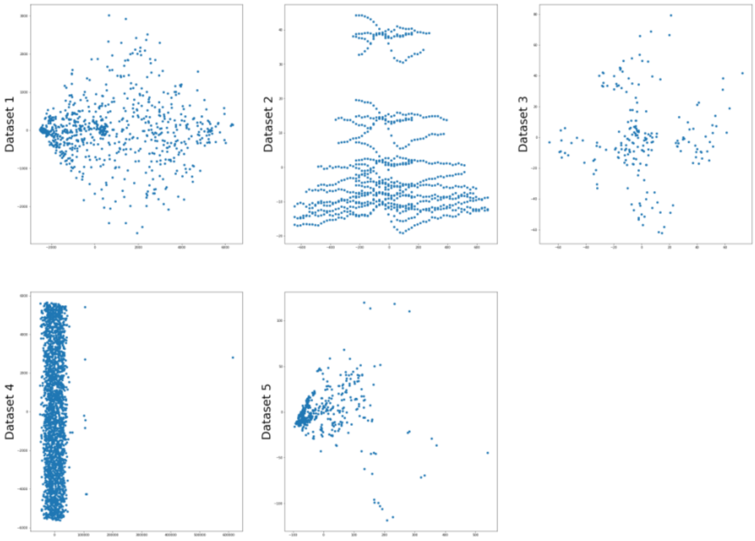


Figure 5.1: Datasets visualization

### 5.1.2 Clustering Validation

Here, performance of two employed techniques studied. Both techniques have been tested from two to hundred clusters. The following Table 5.1 documents the optimal amount of clusters found for each dataset, by the K-Means technique. This is the quantities of clusters where the silhouette index has been computed to be the highest.

Table 5.1: K-Means Performance

| | OPTIMAL AMOUNT OF CLUSTERS FOUND BY K MEANS |
|---|---|
| **D1** | 2 |
| **D2** | 100 |
| **D3** | 5 |
| **D4** | 3 |
| **D5** | 2 |

Figure 5.2 depicts the silhouette coefficient versus the quantities of clusters relation, from the K-Means.

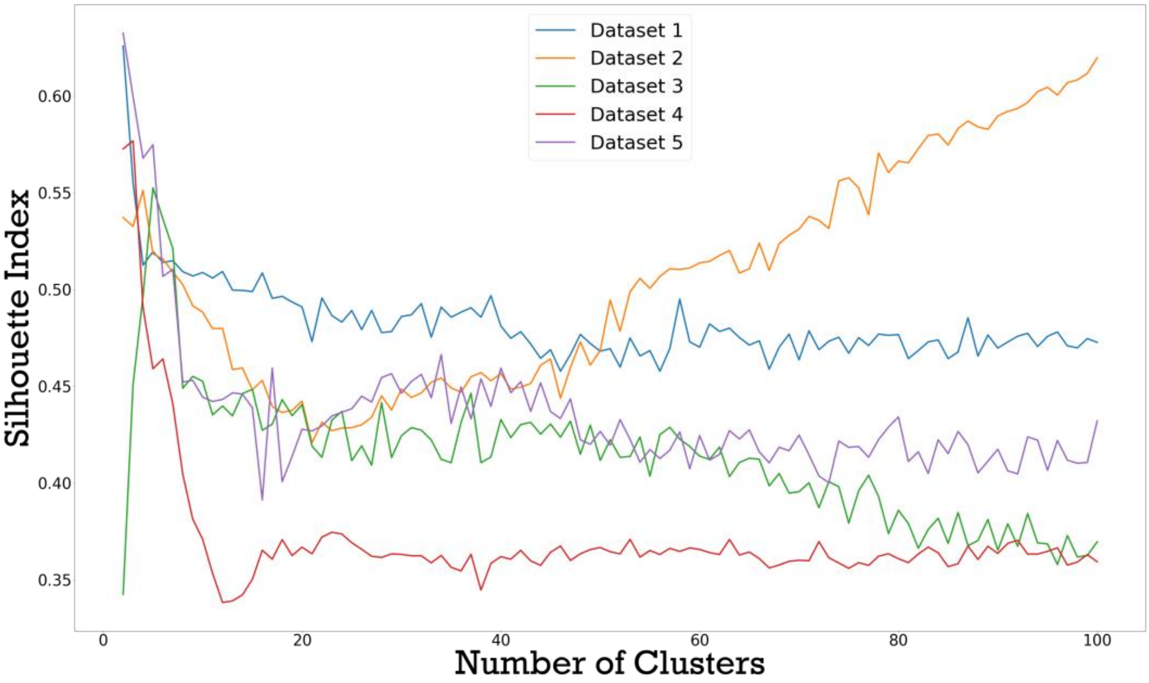


Figure 5.2: Silhouette index found by K-Means

Table 5.2 documents the ideal quantity of clusters found for every dataset, by BIRCH technique.

Table 5.2: BIRCH Performance

| | OPTIMAL AMOUNT OF CLUSTERS FOUND BY BIRCH |
|---|---|
| **D1** | 2 |
| **D2** | 100 |
| **D3** | 5 |
| **D4** | 3 |
| **D5** | 2 |

Figure 5.3 depicts the silhouette coefficient versus the quantity of clusters relation in the five data sets by BIRCH Algorithm.

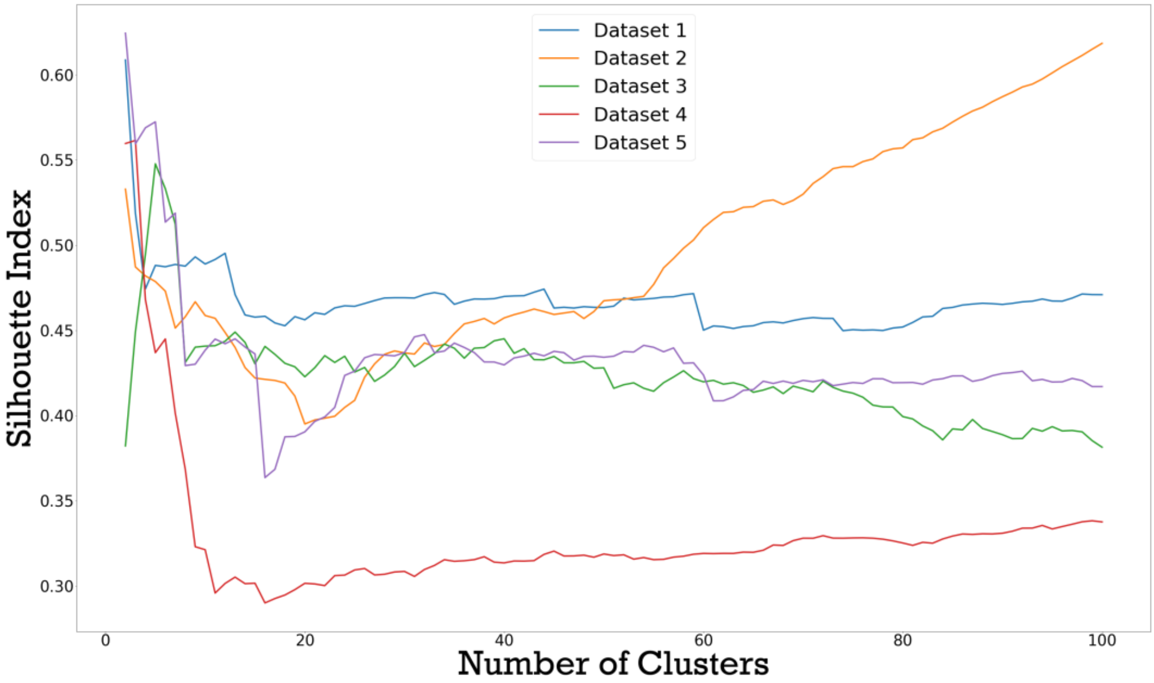


Figure 5.3: Silhouette index found by BIRCH

### 5.1.3 Comparative Analysis

Now, after the Validation has been done, the comparison is done to deduce the comparative results of our article. Table 5.3 documents the analysis in a comparative sense of silhouette coefficient obtained in all the datasets.

Table 5.3: Comparison of K-Means and BIRCH

| | **K-Means** | **Birch** |
|---|---|---|
| **D1** | 0.625829 | 0.608685 |
| **D2** | 0.61976 | 0.61847 |
| **D3** | 0.552626 | 0.547779 |
| **D4** | 0.576879 | 0.561479 |
| **D5** | 0.632506 | 0.624427 |

It is observed, that on all five datasets, K-Means clustering method obtains better result, i.e. a better value for the coefficient. By this observation it is deduced that K means proves out to be a better performing technique in our study.

## 5.2 RESULTS: ANALYSIS & DETECTION OF COVID 19

### Dataset Description

We have gathered CXR pictures from 2 publicly available open-source repositories. In the first step, 616 COVID-19 positive images are collected from GitHub repository maintained by (J. P. Cohen, P. Morrison, 2020) [152]. For the next step, we acquired equal no. of normal & other non-COVID19 pneumonia photos from the Kaggle repository (Mooney, 2018) [153] to prevent any imbalance in the curated dataset between different classes. As a result of this, a final dataset has a total of 1848 X-ray pictures that represent all 3 classes of pneumonia: healthy, COVID-19, & also other non-COVID-19.

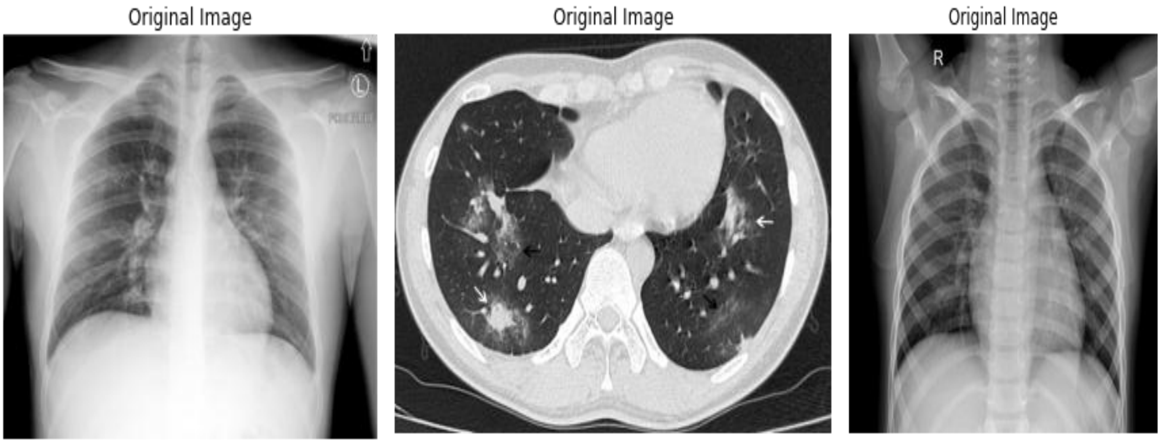


Figure 5.4: Sample Images of collected data

Figure 5.4 shows the samples of dataset images. All sample images were used for resizing, shaping, and segmentation. A variety of clinical contexts have resulted in differing sizes for the CXR pictures in the curated collection. For the input photos to be appropriate for model training and testing, we have done several essential pre-processing activities on them. To achieve this resolution, all photos are reduced in size to 196 x 196 pixels, and indeed the values of all pixels are scaled to the range [0, 1] using the min-max normalization method. Figure 5.4 shows a variety of CXR picture examples from several categories.

Table 5.4: Dataset Labelling

| **Category** | **Label** |
|---|---|
| **Pneumonia** | 0 |
| **Normal** | 1 |
| **Covid** | 2 |

Table 5.4 shows the label and categories of the dataset. The above table shows the data labelling. The Pneumonia category data is denoted by 0, Normal category data is denoted by 1 COVID, and category data is denoted by 2.

## 5.3 SCREENSHOTS OF THE RESULTS

This section visualizes the screenshots of all results obtained from the simulation.

### 1) After Pre-processing Results

Data pre-processing in ML continues with the division of the dataset in the following phase. To prepare a dataset for an ML model, it is necessary to divide it into 2 different sets: training set & testing set. Generally speaking, a training set refers to a subset of a dataset that is used to train an ML model.

Even though chest X-ray pictures always include black, white, and grey components, chest X-ray images always exhibit restricted contrast as a result of the low exposure dosage administered to patients. Due to the location of the lungs on both sides of the chest cavity, the region around the lungs might be easily ignored by X-rays, which are almost black. It looks almost white in this picture because the X-rays cannot travel through the heart, which is located in the middle of the lungs, and so cannot completely penetrate it. Bones are made of protein; therefore since they are too thick to allow X-rays to flow through them, the bones appear almost completely white in the pictures taken with a radiograph. Aside from that, bones have distinct edges.

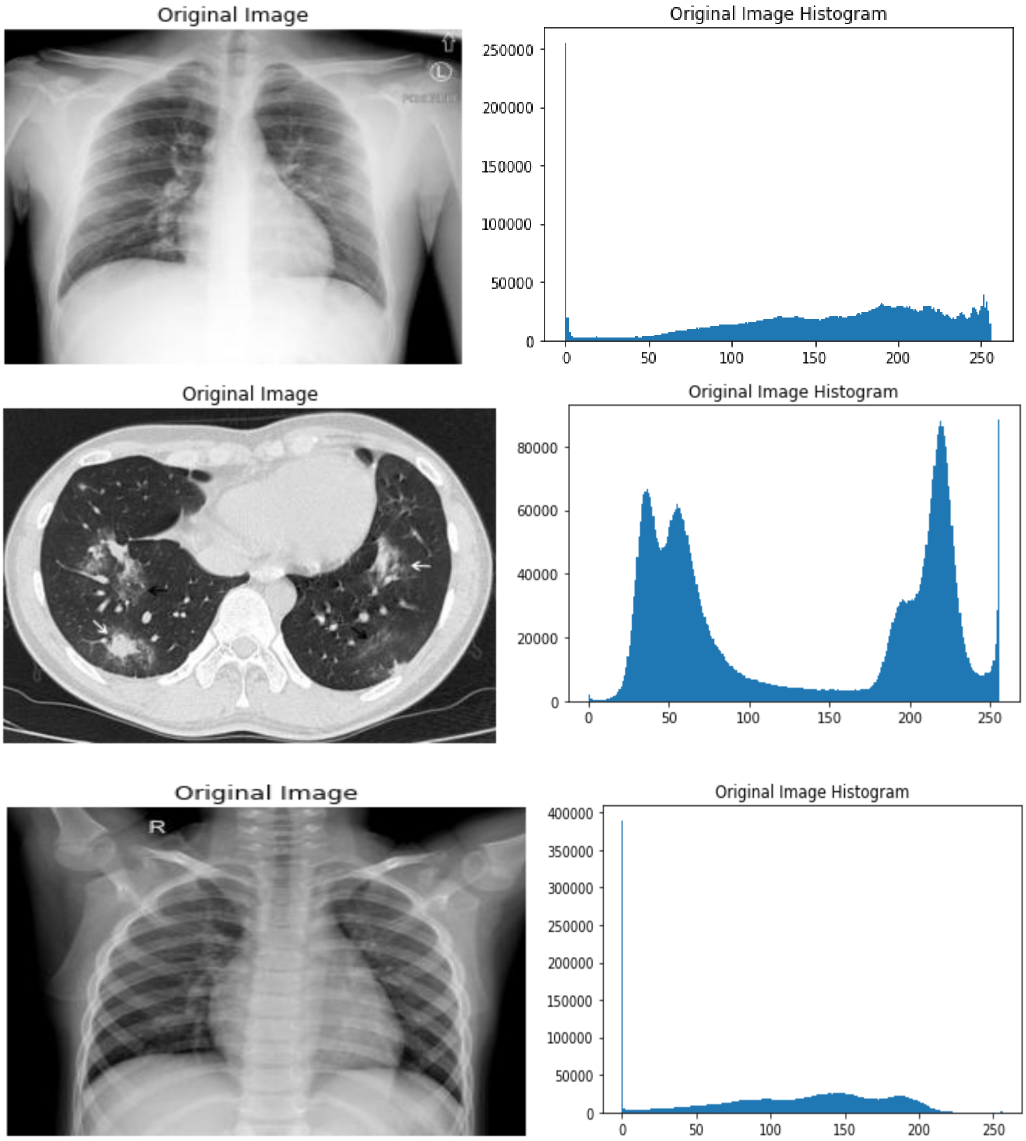


Fig 5.5: Original images with their Histograms

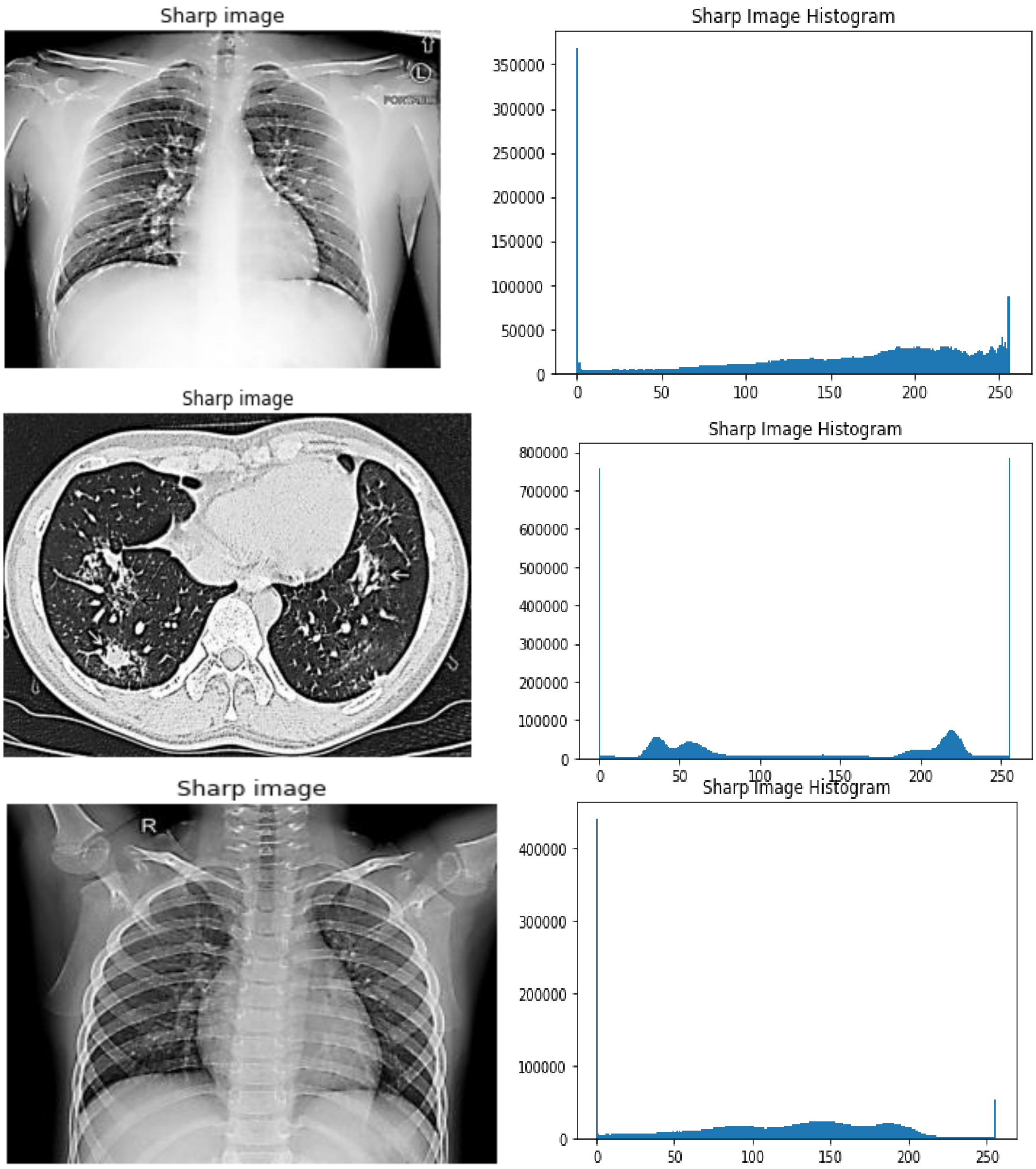


Fig 5.6: Sharp images with their Histograms

It is a graphical representation of the number of pixels in a picture as a function of their intensity that is used to create sharp images using histograms. Each bin in a histogram represents a certain intensity value range, and histograms are composed of bins.

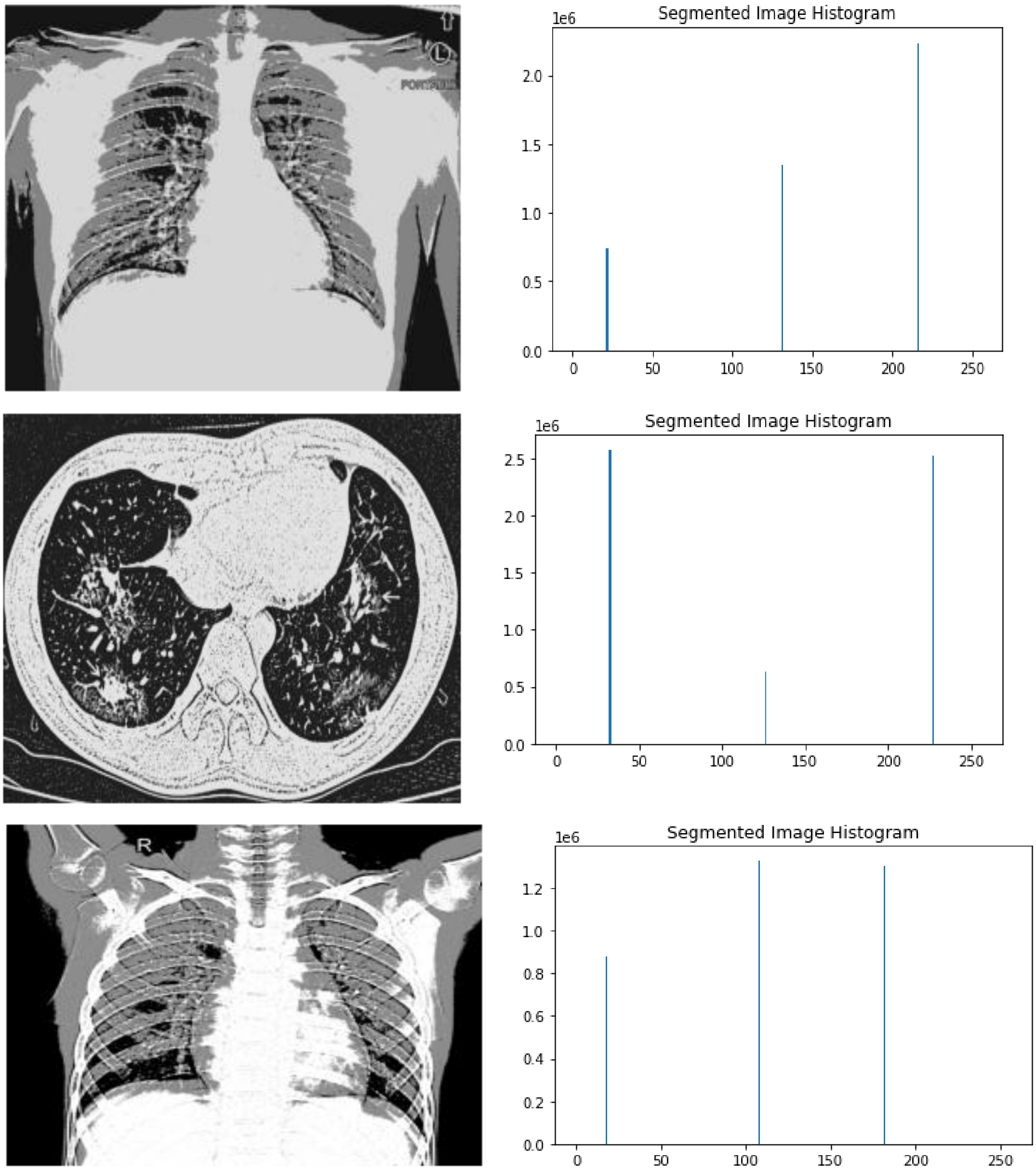


Fig 5.7: Segmented images with their Histograms

The above Figure 5.5, 5.6, 5.7 shows the after pre-processing results with their histograms. In statistics, a histogram is a graphical representation that splits a set of data points into ranges that are determined by the user, and it is used to visualize the distribution of data points. The histogram, which has a similar appearance to a bar graph, consolidates a data series into visually intelligible visuals by gathering multiple data points & separating them into logical ranges or bins.

## 5.4 CLASSIFICATION EVALUATION METRICS

In this subsection, several evaluation metrics, precision, accuracy, recall, F1 score & so on, are described. According to the outputs of the model, four indices, TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative), are utilized to analyze and identify the performance of model. The True Positive means that the CXR images, which suffer from pneumonia, are signed as pneumonia as well by the model. The True Negative means if the chest X-ray images do not show pneumonia & model predicts. The remaining matrices have a similar definition. The four metrics are given as follows:

$$Accuracy = \frac{TP + TN}{\mathrm{TP + TN + FP + FN}} \qquad .... (5.1)$$

$$\mathrm{precision} = \frac{TP}{\mathrm{TP + FP}} \qquad ... (5.2)$$

$$\mathrm{recall} = \frac{TP}{\mathrm{FP + FN}} \qquad ... (5.3)$$

$$\mathrm{F1} = 2 \times \frac{\mathrm{precision}}{\mathrm{precision + recall}} \qquad ..... (5.4)$$

The performance measure of different machine learning algorithms is analyzed by considering measures. Among four metrics, the precision rate was always used to estimate how much the number of images that are truly pneumonia accounted for in the total number examples, which are classified as positive for pneumonia. That is, the pneumonia images must be identified in practical clinical diagnoses and hence, the precision rate is especially important. In most cases, the higher the precision rate gets, the lower the recall rate is. Thus, the F1 score rate is widely considered as a proper criterion.

## 5.5 EXPERIMENTAL RESULTS

During the experimental tests that were conducted, each of the metrics that were included in the practicable also and the proposed solution was examined in detail and evaluated. In addition, the experimental assessment technique is discussed, and the results of the research are provided.

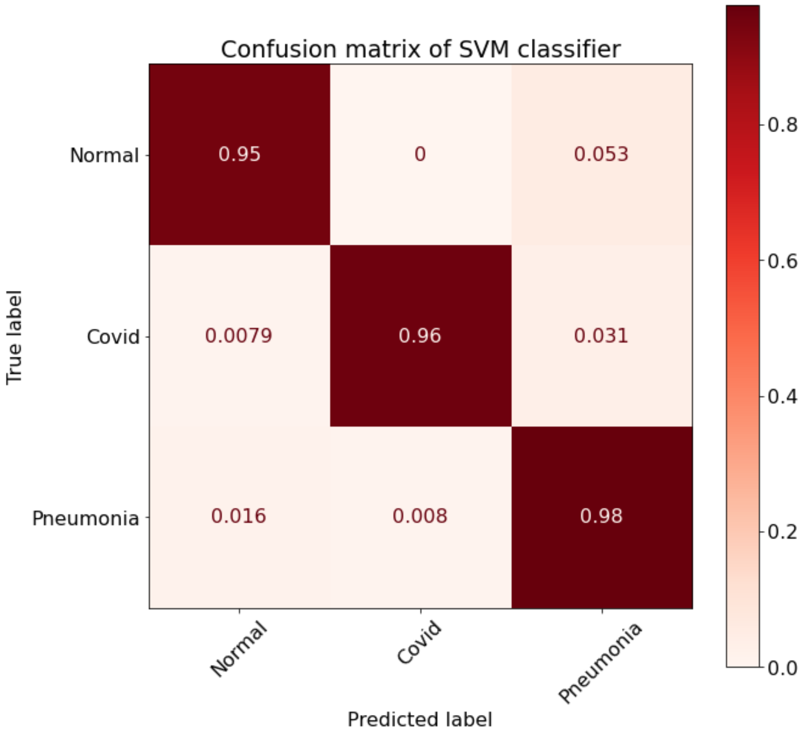


Figure 5.8: Confusion Matrix of SVM Classifier

Figure 5.8 demonstrates confusion matrix of the SVM classifier. The figure depicts the confusion matrix created from the entire system's performance metrics during the classification step. The planned technique requires labelled test data to confirm its expected output during assessment. This depicts complete performance of system. The X-axis shows the predicated label of each dataset category. Similarly, Y-axis shows the true label of three categories which are denoted by the 0, 1 and 2. Based on the confusion matrices and tables, it is clear that the concatenated network outperforms the single-layer network in terms of identifying COVID-19 and not detecting false instances of COVID-19, as well as in terms of outputting higher overall accuracy. Even though we had an imbalanced dataset and just a few examples of COVID-19, we were able to increase COVID-19 detection, as well as the detection of the other classes, by using the suggested approach.

Table 5.5: Results of training and testing accuracy

| Model | Training accuracy | Testing accuracy |
|---|---|---|
| SVM (RBF) | 100 | 96.17 |

Table 5.5 shows performance of training and testing accuracy of our proposed model. M. Shorfuzzaman et al. [66] have carried out classification on same dataset and obtained maximum accuracy of 95.49% on ResNet50V2 model. We have been able to improve the testing accuracy. We are performing Grid Searching for Model, and to find out the best Hyper Parameters. Grid Searching is basically the evaluation of the accuracies of model with different combination of hyperparameters to deduce the best hyperparameters. We have performed this study in python, and after applying this technique we have tuned the hyperparameters. We evaluated the performance of classifier on different combinations of SVM kernels and regularization parameters. After application of this technique, we deduce best kernel for this study is the RBF Kernel, and the Regularization Parameter is 10. The SVM classifier achieved 100% training accuracy and 96.17% testing accuracy. Training accuracy refers to the accuracy with which identical pictures are used for both training & testing, while test accuracy refers to the accuracy with which the trained model detects independent images that were not used in training (or vice versa).

Table 5.6: Results of Different Parameters

| **Parameters** | **Values obtained** |
|---|---|
| Precision | 0.9630 |
| Recall | 0.9620 |
| F1 Score | 0.9620 |

The above Table 5.6 shows the three parameters that are F1 score, Recall, and precision. The F1 score achieved a higher value in comparison to precision and recall. F1 score for SVM classifier is 0.9620 Precision is 0.9630 or Recall is 0.9620.

# CHAPTER

# CONCLUSION AND FUTURE SCOPE

## 6.1 CONCLUSION

With increasing number of data & ML centric studies in industrial applications and academia, clustering is being heavily used in multiple domains for various purposes. Clustering is being heavily used in customer base identification, image processing, and recommendation systems, in libraries to cluster different books on the basis of genre, topic etc. and many more. It is observed that K means and BIRCH have extensively been used in medical domain, customer domain etc. and have proved to be popular and good performing algorithms. These two algorithms are hence taken up in this study to be comparatively evaluated. These two techniques were applied to five datasets. Before the application of clustering techniques, the preprocessing and dimensionality reduction of the datasets was done. The validation of the clustering results was done using the silhouette coefficient. The ideal amounts of clusters for all five datasets were deduced, and the results from the silhouette coefficient were compared. From our results, it is concluded that K-Means obtains higher silhouette index on all five datasets, thus, proves to be a better performing technique. In the mentioned previous studies, comparison of clustering algorithms has been done before too. In previous literature CLARA algorithm has been proven to better than K-Means on certain data, and comparison of K-Means and FCM have also been performed. Along the same lines we have also conducted the first part of the study. As far as the application of K-Means algorithm for detection of COVID is concerned, the diagnosis of SARS CoV-2, that is responsible for COVID-19, utilizing chest X-ray pictures is critical for both patients and clinicians since it may save their lives if they are diagnosed early. As a bonus, in nations where it is not possible to obtain laboratory testing equipment, this becomes even more critical. Because of the huge no. of fatalities caused by the coronavirus pandemic, healthcare systems around the globe have been strained to their breaking point. A more rapid, simpler, and less expensive method of detecting the COVID-19 virus may help save lives & lessen load on healthcare professionals by detecting the virus early. By applying IP methods to X-ray images, machine learning has played a significant role in identification of COVID-19. Furthermore, KMC was used for IS. By merging the characteristics collected by the PCA approach, this research work intended & applied an

ML system for the identification of COVID-19 with high accuracy & minimal complexity. SVM classifier classification technique is crucial in the detection of COVID-19 using CXR images, which was used in this study. It was possible to detect COVID-19 by employing X-ray scans of chest to create output of highlighted lung vital region, which was then used to detect COVID-19. Suggested SVM classifier outperformed competition in terms of classification accuracy.

## 6.2 FUTURE WORK

The work done in future to compare and evaluate the results on the considered algorithms will be critical to our study as it will provide more useful information with respect to our study. Previous studies have taken many algorithms like FCM, CLARA, and other algorithms are also mentioned in the article. The Comparison of these types of clustering algorithms on the datasets we have taken up would be considered an extension to our study. The comparison of the two algorithms we have taken i.e. K-Means and BIRCH on more datasets will also be extension to our work. Application of both algorithms on a real life case study, to observe which algorithm performs better would also be a potential future work. Coming to the application part, the work that will be done in the future to develop, host, and benchmark COVID-19-related data sets will be critical since it will aid in the identification of discoveries that will be valuable in the fight against the illness. The collection of further COVID-19 photos as they become available soon will be an immediate extension of our present effort, to significantly improve prediction findings. Furthermore, we want to use an enhanced fusion model depending upon feature concatenation taken from CXR pictures & multimodal COVID-19 data in the future to increase the accuracy of predictions.

# REFERENCES

[1]. Nur-a-alam, Ahsan, M., Based, M. A., Haider, J., & Kowalski, M. (2021). COVID-19 detection from chest X-ray images using feature fusion and deep learning. Sensors. https://doi.org/10.3390/s21041480

[2]. Holshue, M. L., DeBolt, C., Lindquist, S., Lofy, K. H., Wiesman, J., Bruce, H., Spitters, C., Ericson, K., Wilkerson, S., Tural, A., Diaz, G., Cohn, A., Fox, L., Patel, A., Gerber, S. I., Kim, L., Tong, S., Lu, X., Lindstrom, S., … Pillai, S. K. (2020). First Case of 2019 Novel Coronavirus in the United States. New England Journal of Medicine. https://doi.org/10.1056/nejmoa2001191

[3]. Zhang, R., Tie, X., Qi, Z., Bevins, N. B., Zhang, C., Griner, D., Song, T. K., Nadig, J. D., Schiebler, M. L., Garrett, J. W., Li, K., Reeder, S. B., & Chen, G. H. (2021). Diagnosis of Coronavirus Disease 2019 Pneumonia by Using Chest Radiography: Value of Artificial Intelligence. In Radiology. https://doi.org/10.1148/RADIOL.2020202944

[4]. Of, C. (2020). Detection of COVID-19 Using Chest X-ray Images By Convolutional Neural Networks.

[5]. Rishabh Raj. (2020). CoviDecode : Detection of COVID-19 from Chest X-Ray images using Convolutional Neural Networks. International Journal for Modern Trends in Science and Technology. https://doi.org/10.46501/ijmtst061283

[6]. M. A. Almahfud, R. Setyawan, C. A. Sari, D. R. I. M. Setiadi, and E. H. Rachmawanto: An Effective MRI Brain Image Segmentation using Joint Clustering (K-Means and Fuzzy C-Means). In: 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). IEEE, Yogyakarta, Indonesia. pp. 11–16 (Nov. 2018). https://doi.org/10.1109/ISRITI.2018.8864326.

[7]. R. M. Prakash, K. Bhuvaneshwari, M. Divya, K. J. Sri, and A. S. Begum: Segmentation of thermal infrared breast images using K-means, FCM and EM algorithms for breast cancer detection. In: 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). IEEE, Coimbatore. pp. 1–4 (Mar. 2017). https://doi.org/10.1109/ICIIECS.2017.8276142.

[8]. Y.-P. Huang, P. Singh, and H.-C. Kuo: A Hybrid Fuzzy Clustering Approach for the Recognition and Visualization of MRI Images of Parkinson’s Disease. In:

IEEE Access, vol. 8. pp. 25041–25051 (2020). https://doi.org/10.1109/ACCESS.2020.2969806.

[9]. IUllah, H. Hussain, I. Ali, and A. Liaquat: Churn Prediction in Banking System using K-Means, LOF, and CBLOF. In: 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE). IEEE, Swat, Pakistan. pp. 1–6 (Jul. 2019). https://doi.org/10.1109/ICECCE47252.2019.8940667.

[10]. H. Du and Y. Li. An Improved BIRCH Clustering Algorithm and Application in Thermal Power. In: 2010 International Conference on Web Information Systems and Mining. IEEE, Sanya, China. pp. 53–56 (Oct. 2010). https://doi.org/10.1109/WISM.2010.123.

[11]. G. Pitolli, L. Aniello, G. Laurenza, L. Querzoni, and R. Baldoni. Malware family identification with BIRCH clustering. In: 2017 International Carnahan Conference on Security Technology (ICCST). IEEE, Madrid. pp. 1–6 (Oct. 2017). https://doi.org/10.1109/CCST.2017.8167802.

[12]. Q. Li et al. BIRCH Algorithm and Wasserstein Distance Metric Based Method for Generating Typical Scenarios of Wind Power Outputs. In: 2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia). IEEE, Chengdu, China. pp. 3640–3644 (May 2019). https://doi.org/10.1109/ISGT-Asia.2019.8881562.

[13]. P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. In: Journal of Computational and Applied Mathematics, vol. 20. pp. 53–65 (Nov. 1987). https://doi.org/10.1016/0377-0427(87)90125-7.

[14]. F. Wang, H.-H. Franco-Penya, J. D. Kelleher, J. Pugh, and R. Ross. An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity. In: Machine Learning and Data Mining in Pattern Recognition. vol. 10358, P. Perner, Ed. Cham: Springer International Publishing. pp. 291–305 (2017). https://doi.org/10.1007/978-3-319-62416-7_21.

[15]. Das, A. K., Ghosh, S., Thunder, S., Dutta, R., Agarwal, S., & Chakrabarti, A. (2021). Automatic COVID-19 detection from X-ray images using ensemble learning with convolutional neural network. Pattern Analysis and Applications. https://doi.org/10.1007/s10044-021-00970-4

[16]. Wang, L., Lin, Z. Q., & Wong, A. (2020). COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from

chest X-ray images. Scientific Reports. https://doi.org/10.1038/s41598-020-76550-z

[17]. Qin, C., Yao, D., Shi, Y., & Song, Z. (2018). Computer-aided detection in chest radiography based on artificial intelligence: A survey. In BioMedical Engineering Online. https://doi.org/10.1186/s12938-018-0544-y

[18]. Aziz, M. N., Purboyo, T. W., & Prasasti, A. L. (2017). A survey on the implementation of image enhancement. International Journal of Applied Engineering Research.

[19]. Image Pre-Processing Techniques for X-Ray Medical Images : a Survey. (2021). 9(1), 1999–2002.

[20]. Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K. G., & Murphy, K. (2021). Deep learning for chest X-ray analysis: A survey. In Medical Image Analysis. https://doi.org/10.1016/j.media.2021.102125

[21]. Parveen, S., & Khan, K. B. (2020). Detection and classification of pneumonia in chest X-ray images by supervised learning. Proceedings - 2020 23rd IEEE International Multi-Topic Conference, INMIC 2020. https://doi.org/10.1109/INMIC50486.2020.9318118

[22]. Alić, B., Gurbeta, L., & Badnjević, A. (2017). Machine learning techniques for classification of diabetes and cardiovascular diseases. 2017 6th Mediterranean Conference on Embedded Computing, MECO 2017 - Including ECYPS 2017, Proceedings. https://doi.org/10.1109/MECO.2017.7977152

[23]. Zhou, H., Tang, J., & Zheng, H. (2015). Machine Learning for Medical Applications. In Scientific World Journal. https://doi.org/10.1155/2015/825267

[24]. Reddy, R. V. K., & Babu, U. R. (2018). A review on classification techniques in machine learning. International Journal of Advance Research in Science and Engineering.

[25]. Hormozi, H., Hormozi, E., & Nohooji, H. R. (2012). The Classification of the Applicable Machine Learning Methods in Robot Manipulators. International Journal of Machine Learning and Computing. https://doi.org/10.7763/ijmlc.2012.v2.189

[26]. Meenakshi, M. (2020). Machine Learning Algorithms and their Real-life Applications: A Survey. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3595299

[27]. J. Han, M. Kamber, and J. Pei. Data Mining: Concepts and Techniques, Third Edition.

[28]. Witten, I. H., Frank, E., Trigg, L., Hall, M., Holmes, G., & Cunningham, S. J. (1999). Weka : Practical Machine Learning Tools and Techniques with Java Implementations. Seminar.

[29]. Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. Machine Learning. https://doi.org/10.1007/bf00153759

[30]. Alawnah, S., & Sagahyroon, A. (2017). Modeling of smartphones' power using neural networks. Eurasip Journal on Embedded Systems. https://doi.org/10.1186/s13639-017-0070-1

[31]. Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation. https://doi.org/10.1162/089976601300014493

[32]. Breiman, L. (2001). Random forests. Machine Learning. https://doi.org/10.1023/A:1010933404324

[33]. Cessie, S. Le, & Houwelingen, J. C. Van. (1992). Ridge Estimators in Logistic Regression. Applied Statistics. https://doi.org/10.2307/2347628

[34]. Freund, Y., & Schapire, R. E. (1996). Experiments with a New Boosting Algorithm. Proceedings of the 13th International Conference on Machine Learning. https://doi.org/10.1.1.133.1040

[35]. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. Neurocomputing. https://doi.org/10.1016/j.neucom.2016.12.038

[36]. Asiri, N., Hussain, M., Al Adel, F., & Alzaidi, N. (2019). Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey. In Artificial Intelligence in Medicine. https://doi.org/10.1016/j.artmed.2019.07.009

[37]. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2017). Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2016.2587640

[38]. Kalaivani, K. S., Uma, S., & Kanimozhiselvi, C. S. (2020). A Review on Feature Extraction Techniques for Sentiment Classification. Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-000126

[39]. Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. Proceedings of 2014 Science and Information Conference, SAI 2014. https://doi.org/10.1109/SAI.2014.6918213

[40]. Fang, Y., Zhang, H., Xie, J., Lin, M., Ying, L., Pang, P., & Ji, W. (2020). Sensitivity of chest CT for COVID-19: Comparison to RT-PCR. In Radiology. https://doi.org/10.1148/radiol.2020200432

[41]. Parveen and A. Singh, "Detection of brain tumor in MRI images, using combination of fuzzy c-means and SVM," 2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN), 2015, pp. 98-102, doi: 10.1109/SPIN.2015.7095308.

[42]. N. R. Bhimte and V. R. Thool, "Diseases Detection of Cotton Leaf Spot Using Image Processing and SVM Classifier," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 340-344, doi: 10.1109/ICCONS.2018.8662906.

[43]. A Hussain and A. Khunteta, "Semantic Segmentation of Brain Tumor from MRI Images and SVM Classification using GLCM Features," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 38-43, doi: 10.1109/ICIRCA48905.2020.9183385.

[44]. A Vysala and Dr. J. Gomes. Evaluating and Validating Cluster Results. In: Computer Science & Information Technology. AIRCC Publishing Corporation (2020). pp. 37–47 (Jul. 2020). https://doi.org/10.5121/csit.2020.100904.

[45]. Y. Tu, Y. Liu, and Z. Li. Online Segmentation Algorithm for Time Series Based on BIRCH Clustering Features. In: 2010 International Conference on Computational Intelligence and Security. IEEE, Nanning, Guangxi, TBD, China. pp. 55–59 (Dec. 2010). https://doi.org/10.1109/CIS.2010.19.

[46]. P. Sudheera, V. R. Sajja, S. D. Kumar, and N. G. Rao. Detection of Dental Plaque using Enhanced K-Means and Silhouette Methods. In: 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT). IEEE, Ramanathapuram, India .p. 5 (January 2017). 2016. https://doi.org/10.1109/ICACCCT.2016.7831702

[47]. AK. Singh, S. Mittal, P. Malhotra, and Y. V. Srivastava. Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means. In: 2020 Fourth International Conference on Computing Methodologies and Communication

(ICCMC). IEEE, Erode, India. pp. 306–310 (Mar. 2020). https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00057.

[48]. A Pugazhenthi and L. S. Kumar. Selection of Optimal Number of Clusters and Centroids for K-means and Fuzzy C-means Clustering: A Review. In: 2020 5th International Conference on Computing, Communication and Security (ICCCS). IEEE, Patna, India. pp. 1–4 (Oct. 2020). https://doi.org/10.1109/ICCCS49678.2020.9276978.

[49]. S. Preetha and R. Rayapeddi. Predicting Customer Churn in the Telecom Industry Using Data Analytics. In: 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT). IEEE, Bangalore, India. pp. 38–43 (Aug. 2018). https://doi.org/10.1109/ICGCIoT.2018.8753096.

[50]. H. Mahi, N. Farhi, K. Labed, and D. Benhamed. The Silhouette Index and the K-Harmonic Means algorithm for Multispectral Satellite Images Clustering. In: 2018 International Conference on Applied Smart Systems (ICASS). IEEE, Medea, Algeria. pp. 1–6 (Nov. 2018). https://doi.org/10.1109/ICASS.2018.8652068.

[51]. S.-H. Jun and S.-J. Lee. Empirical Comparisons of Clustering Algorithms using Silhouette Information. In: International Journal of Fuzzy Logic and Intelligent Systems, vol. 10, no. 1. pp. 31–36 (Mar. 2010). https://doi.org/10.5391/IJFIS.2010.10.1.031.

[52]. T. Gupta and S. P. Panda. Clustering Validation of CLARA and K-Means Using Silhouette & DUNN Measures on Iris Dataset. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). IEEE, Faridabad, India. pp. 10–13 (Feb. 2019). https://doi.org/10.1109/COMITCon.2019.8862199.

[53]. AD. Fontanini and J. Abreu. A Data-Driven BIRCH Clustering Method for Extracting Typical Load Profiles for Big Data. In: 2018 IEEE Power & Energy Society General Meeting (PESGM). IEEE, Portland, OR, USA. pp. 1–5 (Aug. 2018). https://doi.org/10.1109/PESGM.2018.8586542.

[54]. M. Aryuni, E. Didik Madyatmadja, and E. Miranda. Customer Segmentation in XYZ Bank Using K-Means and K-Medoids Clustering. In: 2018 International Conference on Information Management and Technology (ICIMTech). IEEE, Jakarta. pp. 412–416 (Sep. 2018). https://doi.org/10.1109/ICIMTech.2018.8528086.

[55]. Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu. Understanding of Internal Clustering Validation Measures. In: 2010 IEEE International Conference on Data Mining. IEEE, Sydney, Australia. pp. 911–916 (Dec. 2010). https://doi.org/10.1109/ICDM.2010.35.

[56]. X. Min and R. Lin. K-Means Algorithm: Fraud Detection Based on Signaling Data. In: 2018 IEEE World Congress on Services (SERVICES). IEEE, San Francisco, CA. pp. 21–22 (Jul. 2018). https://doi.org/10.1109/SERVICES.2018.00024.

[57]. G. V. Bhalerao and N. Sampathila. K-means clustering approach for segmentation of corpus callosum from brain magnetic resonance images. In: International Conference on Circuits, Communication, Control and Computing. IEEE, Bangalore, India. pp. 434–437 (Nov. 2014). https://doi.org/10.1109/CIMCA.2014.7057839.

[58]. R. Alzu'bi, A. Anushya, E. Hamed, B. S. Angela Vincy, and A. AlSha'ar. Medical Image Segmentation via Optimized K-Means. In: 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC). IEEE, Mysore, India. pp. 959–962 (Sep. 2017). https://doi.org/10.1109/CTCEEC.2017.8455030.

[59]. S. Songma, W. Chimphlee, K. Maichalernnukul, and P. Sanguansat. Classification via k-Means Clustering and Distance-Based Outlier Detection. In: 2012 Tenth International Conference on ICT and Knowledge Engineering. IEEE, Bangkok, Thailand. p. 4 (January 2013). https://doi.org/10.1109/ICTKE.2012.6408540.

[60]. T.H. Sardar and Z. Ansari. An Analysis of Distributed Document Clustering Using MapReduce Based K-Means Algorithm. In: Journal of The Institution of Engineers (India): Series B 101. pp.641–650 (2020). https://doi.org/10.1007/s40031-020-00485-2.

[61]. Joy Iong Zong Chen. Automatic Vehicle License Plate Detection using K-Means Clustering Algorithm and CNN. In: Journal of Electrical Engineering and Automation 3, no. 1. pp.15-23 (2021). https://doi.org/10.36548/jeea.2021.1.002.

[62]. K. Zhou and S. Yang. Effect of cluster size distribution on clustering: a comparative study of k-means and fuzzy c-means clustering. In: Pattern Analysis and Applications 23. pp.455–466 (2020). https://doi.org/10.1007/s10044-019-00783-6

[63]. C. Cheng, C. Peng and T. Zhang. Fuzzy K-Means Cluster Based Generalized Predictive Control of Ultra Supercritical Power Plant. In: IEEE Transactions on Industrial Informatics, vol. 17, no. 7. pp. 4575-4583 (July 2021). https://doi.org/10.1109/TII.2020.3020259.

[64]. S. Soor, A. Challa, S. Danda, B. S. D. Sagar and L. Najman. Iterated Watersheds, A Connected Variation of K-Means for Clustering GIS Data. In: IEEE Transactions on Emerging Topics in Computing, vol. 9, no. 2. pp. 626-636 (1 April-June 2021). https://doi.org/10.1109/TETC.2019.2910147.

[65]. I Khan, Z. Luo, J. Z. Huang and W. Shahzad. Variable Weighting in Fuzzy k-Means Clustering to Determine the Number of Clusters. In: IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 9. pp. 1838-1853 (1 Sept. 2020). https://doi.org/10.1109/TKDE.2019.2911582.

[66]. S. De, S. Rakshit, A. Biswas, S. Saha, S. Datta. Implementation of Real-Time Skin Segmentation Based on K-Means Clustering Method. In: Computational Vision and Bio-Inspired Computing, ICCVBIC 2019, Advances in Intelligent Systems and Computing, vol 1108. pp. 964-973 (2019). https://doi.org/10.1007/978-3-030-37218-7_102.

[67]. Shorfuzzaman, M., Masud, M., Alhumyani, H., Anand, D., & Singh, A. (2021). Artificial Neural Network-Based Deep Learning Model for COVID-19 Patient Detection Using X-Ray Chest Images. Journal of Healthcare Engineering. https://doi.org/10.1155/2021/5513679

[68]. Singh, K. K., & Singh, A. (2021). Diagnosis of COVID-19 from chest X-ray images using wavelets-based depthwise convolution network. Big Data Mining and Analytics. https://doi.org/10.26599/BDMA.2020.9020012

[69]. Krishnan, K. S., & Krishnan, K. S. (2021). Vision Transformer based COVID-19 Detection using Chest X-rays. https://doi.org/10.1109/ispcc53510.2021.9609375

[70]. Ji, D., Zhang, Z., Zhao, Y., & Zhao, Q. (2021). Research on Classification of COVID-19 Chest X-Ray Image Modal Feature Fusion Based on Deep Learning. Journal of Healthcare Engineering. https://doi.org/10.1155/2021/6799202

[71]. Peng, Y., Tang, Y., Lee, S., Zhu, Y., Summers, R. M., & Lu, Z. (2021). COVID-19-CT-CXR: A freely accessible and weakly labeled chest X-Ray and CT image collection on COVID-19 from biomedical literature. IEEE Transactions on Big Data. https://doi.org/10.1109/TBDATA.2020.3035935

[72]. Hou, J., & Gao, T. (2021). Explainable DCNN based chest X-ray image analysis and classification for COVID-19 pneumonia detection. Scientific Reports. https://doi.org/10.1038/s41598-021-95680-6

[73]. Hastuti, E. T., Bustamam, A., Anki, P., Amalia, R., & Salma, A. (2021). Performance of True Transfer Learning using CNN DenseNet121 for COVID-19 Detection from Chest X-Ray Images. InHeNce 2021 - 2021 IEEE International Conference on Health, Instrumentation and Measurement, and Natural Sciences. https://doi.org/10.1109/InHeNce52833.2021.9537261

[74]. Madaan, V., Roy, A., Gupta, C., Agrawal, P., Sharma, A., Bologa, C., & Prodan, R. (2021). XCOVNet: Chest X-ray Image Classification for COVID-19 Early Detection Using Convolutional Neural Networks. New Generation Computing. https://doi.org/10.1007/s00354-021-00121-7

[75]. Calderon-Ramirez, S., Yang, S., Moemeni, A., Colreavy-Donnelly, S., Elizondo, D. A., Oala, L., Rodríguez-Capitán, J., Jiménez-Navarro, M., Lopez-Rubio, E., & Molina-Cabello, M. A. (2021). Improving Uncertainty Estimation with Semi-Supervised Deep Learning for COVID-19 Detection Using Chest X-Ray Images. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3085418

[76]. Awasthi, N., Dayal, A., Cenkeramaddi, L. R., & Yalavarthy, P. K. (2021). Mini-COVIDNet: Efficient Lightweight Deep Neural Network for Ultrasound Based Point-of-Care Detection of COVID-19. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control. https://doi.org/10.1109/TUFFC.2021.3068190

[77]. Panetta, K., Sanghavi, F., Agaian, S., & Madan, N. (2021). Automated Detection of COVID-19 Cases on Radiographs using Shape-Dependent Fibonacci-p Patterns. IEEE Journal of Biomedical and Health Informatics. https://doi.org/10.1109/JBHI.2021.3069798

[78]. Channa, A., Popescu, N., & Malik, N. ur R. (2020). Robust Technique to Detect COVID-19 using Chest X-ray Images. https://doi.org/10.1109/ehb50910.2020.9280216

[79]. Liu, J., Zhang, Z., Zu, L., Wang, H., & Zhong, Y. (2020). Intelligent Detection for CT Image of COVID-19 using Deep Learning. Proceedings - 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2020. https://doi.org/10.1109/CISP-BMEI51763.2020.9263690

[80]. Haritha, D., Pranathi, M. K., & Reethika, M. (2020). COVID Detection from Chest X-rays with DeepLearning: CheXNet. Proceedings of the 2020 International Conference on Computing, Communication and Security, ICCCS 2020. https://doi.org/10.1109/ICCCS49678.2020.9277077

[81]. Ahsan Pritom, M. M., Schweitzer, K. M., Bateman, R. M., Xu, M., & Xu, S. (2020). Data-Driven Characterization and Detection of COVID-19 Themed Malicious Websites. Proceedings - 2020 IEEE International Conference on Intelligence and Security Informatics, ISI 2020. https://doi.org/10.1109/ISI49825.2020.9280522

[82]. Kapetanović, A. L., & Poljak, D. (2020). Modeling the Epidemic Outbreak and Dynamics of COVID-19 in Croatia. 2020 5th International Conference on Smart and Sustainable Technologies (SpliTech), 1–5. https://doi.org/10.23919/SpliTech49282.2020.9243757

[83]. Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M. K., Zhang, Y., Sun, L., & Wang, G. (2018). Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss. IEEE Transactions on Medical Imaging. https://doi.org/10.1109/TMI.2018.2827462

[84]. Van Tulder, G., & De Bruijne, M. (2016). Combining Generative and Discriminative Representation Learning for Lung CT Analysis With Convolutional Restricted Boltzmann Machines. IEEE Transactions on Medical Imaging. https://doi.org/10.1109/TMI.2016.2526687

[85]. Dimas, S. R. D., Negara, B. S., Sanjaya, S., & Satria, E. (2021). COVID-19 Classification for Chest X-Ray Images using Deep Learning and Resnet-101. 2021 International Congress of Advanced Technology and Engineering, ICOTEN 2021. https://doi.org/10.1109/ICOTEN52080.2021.9493431

[86]. Rawat, R. M., Garg, S., Jain, N., & Gupta, G. (2021). COVID-19 detection using convolutional neural network architectures based upon chest X-rays images. Proceedings - 5th International Conference on Intelligent Computing and Control Systems, ICICCS 2021. https://doi.org/10.1109/ICICCS51141.2021.9432134

[87]. Akter, S., Shamrat, F. M. J. M., Chakraborty, S., Karim, A., & Azam, S. (2021). Covid-19 detection using deep learning algorithm on chest X-ray images. Biology. https://doi.org/10.3390/biology10111174

[88]. Shadin, N. S., Sanjana, S., & Lisa, N. J. (2021). COVID-19 Diagnosis from Chest X-ray Images Using Convolutional Neural Network(CNN) and InceptionV3. 2021 International Conference on Information Technology, ICIT 2021 - Proceedings. https://doi.org/10.1109/ICIT52682.2021.9491752

[89]. Wu, Y. H., Gao, S. H., Mei, J., Xu, J., Fan, D. P., Zhang, R. G., & Cheng, M. M. (2021). JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation. IEEE Transactions on Image Processing. https://doi.org/10.1109/TIP.2021.3058783

[90]. Chang, Y.-C., Liu, A.-S., & Chu, W.-C. (2021). Using Deep Learning Algorithms in Chest X-ray Image COVID-19 Diagnosis. https://doi.org/10.1109/ecbios51820.2021.9510393

[91]. Morís, D. I., de Moura, J., Novo, J., & Ortega, M. (2021). Cycle generative adversarial network approaches to produce novel portable chest X-rays images for covid-19 diagnosis. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. https://doi.org/10.1109/ICASSP39728.2021.9414031

[92]. Lin, Z., He, Z., Xie, S., Wang, X., Tan, J., Lu, J., & Tan, B. (2021). AANet: Adaptive Attention Network for COVID-19 Detection from Chest X-Ray Images. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2021.3114747

[93]. Fang, Z., Ren, J., MacLellan, C., Li, H., Zhao, H., Hussain, A., & Fortino, G. (2021). A Novel Multi-stage Residual Feature Fusion Network for Detection of COVID-19 in Chest X-ray Images. IEEE Transactions on Molecular, Biological, and Multi-Scale Communications. https://doi.org/10.1109/TMBMC.2021.3099367

[94]. Arias-Londono, J. D., Gomez-Garcia, J. A., Moro-Velazquez, L., & Godino-Llorente, J. I. (2020). Artificial Intelligence applied to chest X-Ray images for the automatic detection of COVID-19. A thoughtful evaluation approach. IEEE Access. https://doi.org/10.1109/ACCESS.2020.3044858

[95]. Sethi, R., Mehrotra, M., & Sethi, D. (2020). Deep Learning based Diagnosis Recommendation for COVID-19 using Chest X-Rays Images. Proceedings of the 2nd International Conference on Inventive Research in Computing Applications, ICIRCA 2020. https://doi.org/10.1109/ICIRCA48905.2020.9183278

[96]. Bekhet, S., Hassaballah, M., Kenk, M. A., & Hameed, M. A. (2020). An Artificial Intelligence Based Technique for COVID-19 Diagnosis from Chest X-Ray. 2nd Novel Intelligent and Leading Emerging Sciences Conference, NILES 2020. https://doi.org/10.1109/NILES50944.2020.9257930

[97]. Jabber, B., Lingampalli, J., Basha, C. Z., & Krishna, A. (2020). Detection of covid-19 patients using chest x-ray images with convolution neural network and mobile net. Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020. https://doi.org/10.1109/ICISS49785.2020.9316100

[98]. Liang, Z., Huang, J. X., Li, J., & Chan, S. (2020). Enhancing Automated COVID-19 Chest X-ray Diagnosis by Image-to-Image GAN Translation. Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020. https://doi.org/10.1109/BIBM49941.2020.9313466

[99]. Calderon-Ramirez, S., Giri, R., Yang, S., Moemeni, A., Umaña, M., Elizondo, D., Torrents-Barrena, J., & Molina-Cabello, M. A. (2020). Dealing with scarce labelled data: Semi-supervised deep learning with mix match for Covid-19 detection using chest X-ray images. Proceedings - International Conference on Pattern Recognition. https://doi.org/10.1109/ICPR48806.2021.9412946

[100]. Al Mamlook, R. E., Chen, S., & Bzizi, H. F. (2020). Investigation of the performance of Machine Learning Classifiers for Pneumonia Detection in Chest X-ray Images. IEEE International Conference on Electro Information Technology. https://doi.org/10.1109/EIT48999.2020.9208232

[101]. Tsai, M. J., & Tao, Y. H. (2019). Machine Learning Based Common Radiologist-Level Pneumonia Detection on Chest X-rays. 2019, 13th International Conference on Signal Processing and Communication Systems, ICSPCS 2019 - Proceedings. https://doi.org/10.1109/ICSPCS47537.2019.9008684

[102]. Katona, T., & Antal, B. (2019). Automated analysis of radiology images using convolutional neural networks. International Symposium on Image and Signal Processing and Analysis, ISPA. https://doi.org/10.1109/ISPA.2019.8868764

[103]. Tang, Y. X., Tang, Y. B., Han, M., Xiao, J., & Summers, R. M. (2019). Abnormal chest x-ray identification with generative adversarial one-class classifier. Proceedings - International Symposium on Biomedical Imaging. https://doi.org/10.1109/ISBI.2019.8759442

[104]. Kieu, P. N., Tran, H. S., Le, T. H., Le, T., & Nguyen, T. T. (2018). Applying Multi-CNNs model for detecting abnormal problem on chest x-ray images. Proceedings of 2018 10th International Conference on Knowledge and Systems Engineering, KSE 2018. https://doi.org/10.1109/KSE.2018.8573404

[105]. Jabra, M. Ben, Koubaa, A., Benjdira, B., Ammar, A., & Hamam, H. (2021). Covid-19 diagnosis in chest x-rays using deep learning and majority voting. Applied Sciences (Switzerland). https://doi.org/10.3390/app11062884

[106]. Hasoon, J. N., Fadel, A. H., Hameed, R. S., Mostafa, S. A., Khalaf, B. A., Mohammed, M. A., & Nedoma, J. (2021). COVID-19 anomaly detection and classification method based on supervised machine learning of chest X-ray images. Results in Physics. https://doi.org/10.1016/j.rinp.2021.105045

[107]. Haque, S., Hoque, M. A., Khan, M. A. I., & Ahmed, S. (2021). COVID-19 Detection Using Feature Extraction and Semi-Supervised Learning from Chest X-ray Images. TENSYMP 2021 - 2021 IEEE Region 10 Symposium. https://doi.org/10.1109/TENSYMP52854.2021.9550977

[108]. Chandra, T. B., Verma, K., Singh, B. K., Jain, D., & Netam, S. S. (2021). Coronavirus disease (COVID-19) detection in Chest X-Ray images using majority voting based classifier ensemble. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2020.113909

[109]. Sheeba Rani, S., Selvakumar, S., Pradeep Mohan Kumar, K., Thanh Tai, D., & Dhiravida Chelvi, E. (2021). 34 - Internet of Medical Things (IoMT) with machine learning–based COVID-19 diagnosis model using chest X-ray images. In U. Kose, D. Gupta, V. H. C. de Albuquerque, & A. Khanna (Eds.), Data Science for COVID-19 (pp. 627–641). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-824536-1.00001-0

[110]. Izdihar, K., Karim, M. K. A., Aresli, N. N., Radzi, S. F. M., Sabarudin, A., Yunus, M. M., Rahman, M. A. A., & Shamsul, S. (2021). Detection of Novel Coronavirus from Chest X-Ray Radiograph Images via Automated Machine Learning and CAD4COVID. 2021 International Congress of Advanced Technology and Engineering, ICOTEN 2021. https://doi.org/10.1109/ICOTEN52080.2021.9493542

[111]. Silva, R. S. R., & Fernando, P. (2021). An extensive survey of machine learning based approaches on automated pathology detection in chest X-rays. Conference

of Open Innovation Association, FRUCT. https://doi.org/10.23919/FRUCT50888.2021.9347605

[112]. Meng, Z., Meng, L., & Tomiyama, H. (2021). Pneumonia Diagnosis on Chest X-Rays with Machine Learning. Procedia Computer Science. https://doi.org/10.1016/j.procs.2021.04.032

[113]. Brunese, L., Martinelli, F., Mercaldo, F., & Santone, A. (2020). Machine learning for coronavirus covid-19 detection from chest x-rays. Procedia Computer Science. https://doi.org/10.1016/j.procs.2020.09.258

[114]. Yee, S. L. K., & Raymond, W. J. K. (2020). Pneumonia Diagnosis Using Chest X-ray Images and Machine Learning. ACM International Conference Proceeding Series. https://doi.org/10.1145/3397391.3397412

[115]. Eljamassi, D. F., & Maghari, A. Y. (2020). COVID-19 Detection from Chest X-ray Scans using Machine Learning. Proceedings - 2020 International Conference on Promising Electronic Technologies, ICPET 2020. https://doi.org/10.1109/ICPET51420.2020.00009

[116]. Thepade, S. D., & Jadhav, K. (2020). Covid19 identification from chest x-ray images using local binary patterns with assorted machine learning classifiers. 2020 IEEE Bombay Section Signature Conference, IBSSC 2020. https://doi.org/10.1109/IBSSC51096.2020.9332158

[117]. Thepade, S. D., Chaudhari, P. R., Dindorkar, M. R., & Bang, S. V. (2020). Covid19 identification using machine learning classifiers with histogram of luminance chroma features of chest x-ray images. 2020 IEEE Bombay Section Signature Conference, IBSSC 2020. https://doi.org/10.1109/IBSSC51096.2020.9332160

[118]. Luo, L., Yu, L., Chen, H., Liu, Q., Wang, X., Xu, J., & Heng, P. A. (2020). Deep Mining External Imperfect Data for Chest X-Ray Disease Screening. IEEE Transactions on Medical Imaging. https://doi.org/10.1109/TMI.2020.3000949

[119]. D. Narin and T. O. Onur, “Investigation of the effect of edge detection algorithms in the detection of Covid-19 patients with convolutional neural network-based features on chest X-ray images,” 2021, doi: 10.1109/SIU53274.2021.9477882.

[120]. S. Lafraxo and M. El Ansari, “CoviNet: Automated COVID-19 detection from X-rays using deep learning techniques,” 2020, doi: 10.1109/CiSt49399.2021.9357250.

[121]. E. Irmak, “A Novel Deep Convolutional Neural Network Model for COVID-19 Disease Detection,” 2020, doi: 10.1109/TIPTEKNO50054.2020.9299286.

[122]. A Bhattacharyya, D. Bhaik, S. Kumar, P. Thakur, R. Sharma, and R. B. Pachori, “A deep learning based approach for automatic detection of COVID-19 cases using chest X-ray images,” Biomed. Signal Process. Control, 2022, doi: 10.1016/j.bspc.2021.103182.

[123]. AH Barshooi and A. Amirkhani, “A novel data augmentation based on Gabor filter and convolutional deep learning for improving the classification of COVID-19 chest X-Ray images,” Biomed. Signal Process. Control, 2022, doi: 10.1016/j.bspc.2021.103326.

[124]. S. Sheykhivand et al., “Developing an efficient deep neural network for automatic detection of COVID-19 using chest X-ray images,” Alexandria Eng. J., 2021, doi: 10.1016/j.aej.2021.01.011.

[125]. K. U. Ahamed et al., “A deep learning approach using effective preprocessing techniques to detect COVID-19 from chest CT-scan and X-ray images,” Comput. Biol. Med., 2021, doi: 10.1016/j.compbiomed.2021.105014.

[126]. N. Rashid, M. A. F. Hossain, M. Ali, M. Islam Sukanya, T. Mahmud, and S. A. Fattah, “AutoCovNet: Unsupervised feature learning using autoencoder and feature merging for detection of COVID-19 from chest X-ray images,” Biocybern. Biomed. Eng., 2021, doi: 10.1016/j.bbe.2021.09.004.

[127]. Halterman, R. L. (2011). Learning to program with Python. In Physiological Research.

[128]. Jackson, C. (2006). Learning to program using Python. Learning to Program Using Python.

[129]. Butwall, M., Ranka, P., & Shah, S. (2019). Python in Field of Data Science: A Review. International Journal of Computer Applications. https://doi.org/10.5120/ijca2019919404

[130]. Beklemysheva, A. (2020). Why Use Python for AI and Machine Learning.

[131]. Jan Erik Solem. (2012). Programming Computer Vision with Python. Programming Computer Vision with Python.

[132]. Santos, B. (2019). An Introduction to Pandas in Python. Towards Data Science.

[133]. Van Der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). Scikit-image: Image processing in python. PeerJ. https://doi.org/10.7717/peerj.453

[134]. Northwestern. (2019). Jupyter Notebook. Northwestern University.

[135]. Software Architectural Styles, https://kaggle.com/qadeemkhan/dataset-of-software-architectural-styles.

[136]. Mall Customer Segmentation Data, https://kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python.

[137]. Marketing Campaign, https://kaggle.com/rodsaldanha/arketing-campaign.

[138]. Birds' Bones and Living Habits, https://kaggle.com/zhangjuefei/birds-bones-and-living-habits.

[139]. Zaafouri, A., Sayadi, M., & Fnaiech, F. (2011). A developed unsharp masking method for images contrast enhancement. International Multi-Conference on Systems, Signals and Devices, SSD'11 - Summary Proceedings. https://doi.org/10.1109/SSD.2011.5767378

[140]. Song, Y., & Yan, H. (2018). Image Segmentation Techniques Overview. AMS 2017 - Asia Modelling Symposium 2017 and 11th International Conference on Mathematical Modelling and Computer Simulation. https://doi.org/10.1109/AMS.2017.24

[141]. Ghoniemy, S. (2016). Medical Image Segmentation Techniques: An Overview. International Journal of Informatics and Medical Data Processing.

[142]. Kim, W., Kanezaki, A., & Tanaka, M. (2020). Unsupervised Learning of Image Segmentation Based on Differentiable Feature Clustering. IEEE Transactions on Image Processing. https://doi.org/10.1109/TIP.2020.3011269

[143]. Liu, H. M., & Lu, J. G. (2015). Brief Survey of K-Means Clustering Algorithms. Applied Mechanics and Materials. https://doi.org/10.4028/www.scientific.net/amm.740.624

[144]. Eamani, R. R., Vinodh Kumar, N., & Jakkamsetti, G. R. (2020). K-means clustering algorithm and architecture: A brief survey. International Journal of Advanced Science and Technology.

[145]. Song, F., Guo, Z., & Mei, D. (2010). Feature selection using principal component analysis. Proceedings - 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization, ICSEM 2010. https://doi.org/10.1109/ICSEM.2010.14

[146]. Kaur, D., & Sharma, S. (2019). Various feature extraction and classification techniques. Lecture Notes in Electrical Engineering. https://doi.org/10.1007/978-981-10-8234-4_51

[147]. Kristiyanti, D. A., & Wahyudi, M. (2017). Feature selection based on Genetic algorithm, particle swarm optimization and principal component analysis for opinion mining cosmetic product review. 2017 5th International Conference on Cyber and IT Service Management, CITSM 2017. https://doi.org/10.1109/CITSM.2017.8089278

[148]. Suganthy, M., & Ramamoorthy, P. (2012). Principal component analysis based feature extraction, morphological edge detection and localization for fast iris recognition. Journal of Computer Science. https://doi.org/10.3844/jcssp.2012.1428.1433

[149]. Erdaw, Y., & Tachbele, E. (2021). Machine learning model applied on chest X-ray images enables automatic detection of COVID-19 cases with high accuracy. International Journal of General Medicine. https://doi.org/10.2147/IJGM.S325609

[150]. Chen, Y., Patel, V. M., Phillips, P. J., Chellappa, R., Poon, T. W. K., Friesen, M. R., Wang, X., Li, X., Leung, V. C. M., Shukla, S., Yadav, R. N., Zorzi, M., Zanella, A., Testolin, A., Grazia, M. D. F. De, Zorzi, M., Guo, L., Jin, B., Yu, R., … Kose, U. (2018). An Optimizing and Differentially Private Clustering Algorithm for Mixed Data in SDN-Based Smart Grid. IEEE Access.

[151]. Mago, N., Srivastava, S., Shirwaikar, R. D., Acharya, U. D., Lewis, L. E. S., & Shivakumar, M. (2016). Prediction of Apnea of Prematurity in neonates using Support Vector Machines and Random Forests. Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016. https://doi.org/10.1109/IC3I.2016.7918051

[152]. J. P. Cohen, P. Morrison, and L. D. (2020). COVID-19 image data collection. Github. https://github.com/ieee8023/covid-chestxray-dataset

[153]. Mooney, P. (2018). Chest X-Ray Images (Pneumonia). Kaggle. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.

# LIST OF PUBLICATIONS

1. Rohan Tomar and Abhilasha Sharma, “K-Means & BIRCH: A Comparative Analysis Study”, 6th International Conference on Inventive Communication and Computational Technologies (ICICCT 2022). (Status: Accepted)
2. R. Tomar and A. Sharma, "Analysis & Detection of COVID-19 on Chest X- Ray Images based on Support Vector Machines," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 2022, pp. 1583-1590, doi: 10.1109/ICSCDS53736.2022.9760748.