

**Data Mining**

***COVID-19 and Pneumonia prediction from medical images***

***Team number***

*Siva Yogitha Mokkapati* sqm6303@psu.edu

*Vineeta Peddinti*  [vkp5111@psu.edu](mailto:vkp5111@psu.edu)

**School of Graduate Professional Studies**

Data Analytics

SWENG 545 – Data Mining

Spring, 2020

# Document Control

## Work carried out by:

|  |  |  |
| --- | --- | --- |
| **Name** | **Email Address** | **Task description** |
| Siva Yogitha Mokkapati | sqm6303@psu.edu | Worked on image processing and clustering techniques |
| Vineeta Peddinti | [vkp5111@psu.edu](mailto:vkp5111@psu.edu) | Worked on other image processing and clustering techniques |

## Revision Sheet

|  |  |
| --- | --- |
| **Date** | **Revision Description** |
| 3/25/2020 | Performed the basic image array conversion and clustering |
| 3/30/2020 | Performed the image array conversion using SIFT and clustering |
| 4/5/2020 | Other clustering techniques |
| 4/10/2020 | Performance metrics for clustering and classification |
| 4/15/2020 | Transfer learning using Inception |
| 4/20/2020 | Got the dataset for COVID |
| 4/25/2020 | Implemented the same for COVID |

**TABLE OF CONTENTS**

**Introduction**

**Problem Statement**

**Importance**

* **Social/commercial/behavioral/scientific/economic)**

**Data**

* **Data Source**
* **Snapshot of the data set**
* **Data Description**

**Preprocessing?**

* **Technique**

**Predictive Analytics technique**

* **Methodology**
* **Why is it chosen?**

**Results**

* **Necessary Figures**
* **Graphs**
* **Tables**

**Discussion**

* **Discussion of results**

**References**

* **Any artworks/ figures you use for the techniques/ information on data/**
* **News or artifacts about the social perspective or implication etc.**
* **Paper / publications/White paper etc.**

**1: Introduction**

Chest X-rays gives images of our heart, lungs, blood vessels, airways, and bones of our chest and spine. It reveals fluid in or around the chest. It’s common way to diagnose disease. The most important thing is that it can also be used to tell whether a certain treatment is working. Generally, people have a series of chest X-rays done over time, to track whether a health problem is getting better or worse. So, in-order to solve the problem of manually classifying X-rays, we are building a model where the images are classified to Normal or Pneumonia, COVID or not. Assuming that the class labels are not previously there, we are turning the classification problem to clustering problem.

**2: Problem Statement**

* To cluster the Chest X-ray images into two classes and differentiate them either as pneumonia condition or non-pneumonia condition using SIFT image matching technique, general image to array and transfer learning using inception to extract the features.
* To repeat the same procedure on the recent CT scan images to differentiate positive and negative for COVID 19.

The overall idea is to understand the ability of image matching techniques and transfer learning techniques for differentiating between images and cluster them.

**2.1: Importance**

Today, we can see how people are suffering from many breathing problems and how late diagnosis and treatment is causing the death of a person. Taking that as our main factor, we want to build models to rightly cluster or classify the chest X-rays.

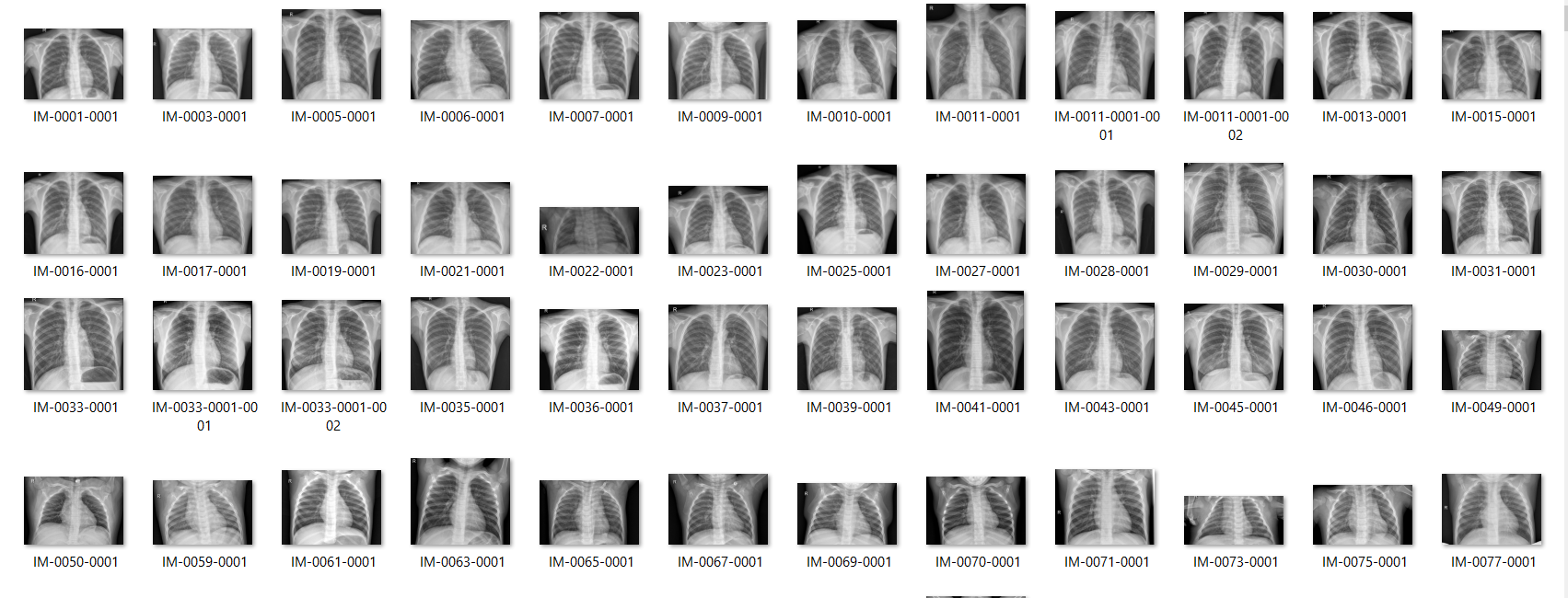
Also, with the increase COVID-19 global pandemic we can see how people are not able to get themselves checked because of the lack of equipment. Currently, the only way to diagnose COVID is either serology or RTPCR. CT-Scans are not able to perform well in diagnosing COVID. Hence, we want to check if CT-scans can be used for the diagnosis.

**3: Data**

* The data source for Pneumonia CXRs is Kaggle.

<https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia>

* The dataset consists of 3875 of pneumonia and non-pneumonia CXR images.
* The data source for COVID CXRs seeded from GitHub COVID-Chest X-ray-dataset. [https](file:///C:\Users\sivay\Downloads\https)[://github.com/ieee8023/covid-chestxray-dataset](https://github.com/ieee8023/covid-chestxray-dataset)
* The dataset is still under development stage and it has a total of 150 images for now. We have added some of the images from Pneumonia and Normal to this dataset in order to balance the classes.



3.1. Snapshot of the data set1

**3.1: Data Preprocessing**

Data is preprocessed according to the techniques applied.

* For image to array conversion, noise is removed from the data and then converted to images. Later the arrays are reshaped to the required format and then normalized.
* For SIFT image matching technique, after converting to array SIFT image matching technique is applied and then the data is reshaped and normalized.
* For transfer learning, images are given to Inception model from which features are extracted which are further reshaped and normalized.

Normalization for images means dividing by 255.

For all of them, after these steps the array contains the features.

**4: Methodology**

We wanted to try traditional image matching technique (SIFT) and cluster the data using that. In addition to that we wanted to compare the traditional Technique with the current state of art for classification which is neural networks. For this purpose, we used the concept of transfer learning in which we used Inception model to extract the image features and used those features to cluster. Also, before the image matching techniques were discovered images were directly converted to array and used as features. Hence, our goal is to understand all the three types and check which is giving better results for our datasets.

After the features were extracted using each technique separately, the image array is given for clustering algorithms. The number of clusters is assigned to 2 as already known.

The clustering algorithms used are: K Means, BIRCH, Hierarchical clustering (single linkage, complete linkage and ward). Since we already have predicted values, we compared them with the actual values and classification report, confusion matrix is generated. Since the classification problem is converted to a clustering problem performance metrics for clustering (within Cluster Sum of Squares, Between Cluster Sum of Squares, CV Score, Silhouette coefficient) are also considered.

Along with these PCA analysis is performed and the top 2 PCA’s are taken. These PC’S are considered as features and clustering is performed using them. The same performance metrics is considered for these features also.

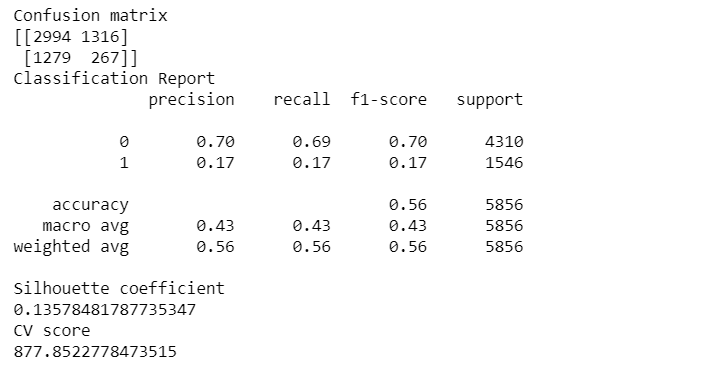
Reason why classification problem is turned into clustering problem:

As stated earlier the goal of the project is to understand the capability of our methods in identifying the appropriate disease from the Chest X Rays (especially COVID). This means that we can have previously unknown diseases, dataset which are not labelled, etc. So, it is better to consider this as a clustering problem and see how well it can identify the small changes in the images.

Python is used as the platform for coding.

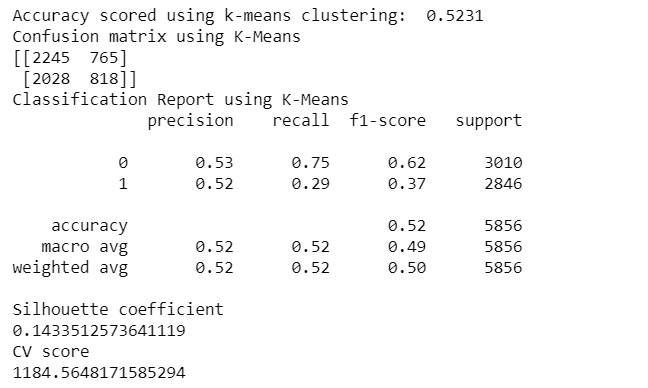
**5: Results**

For **Pneumonia** with traditional method almost all the clustering algorithms gave accuracy between 50-55. Among them the best result was given by ward (56% accuracy). The performance metrics for that is as follows:

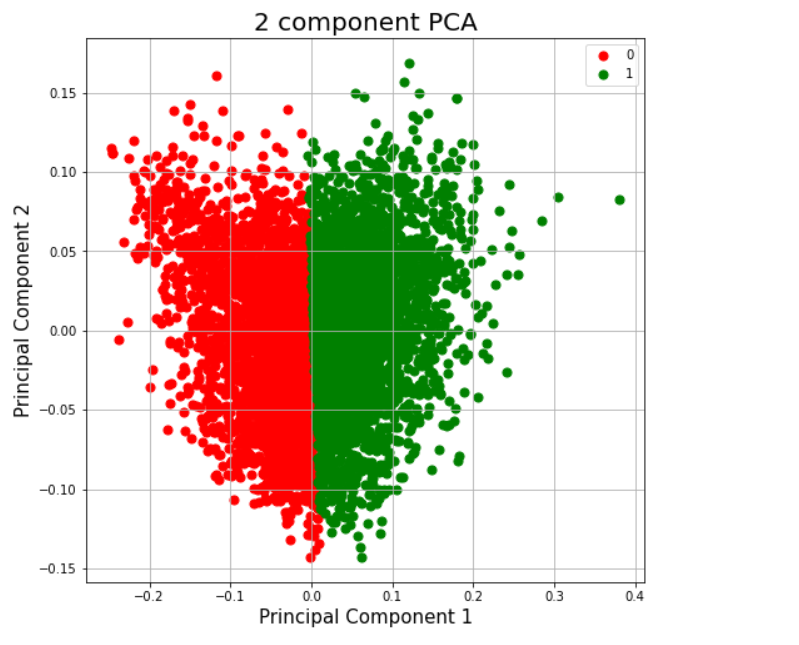


5.1. Performance for ward on pneumonia using traditional method

The first 2 PC components explained approximately 38% of the original data, which is not very good. Hence, performing PCA for this is not very helpful. However, PCA is performed and the accuracy was around 50% for all the algorithms among which K-means is the best and the performance metrics is:

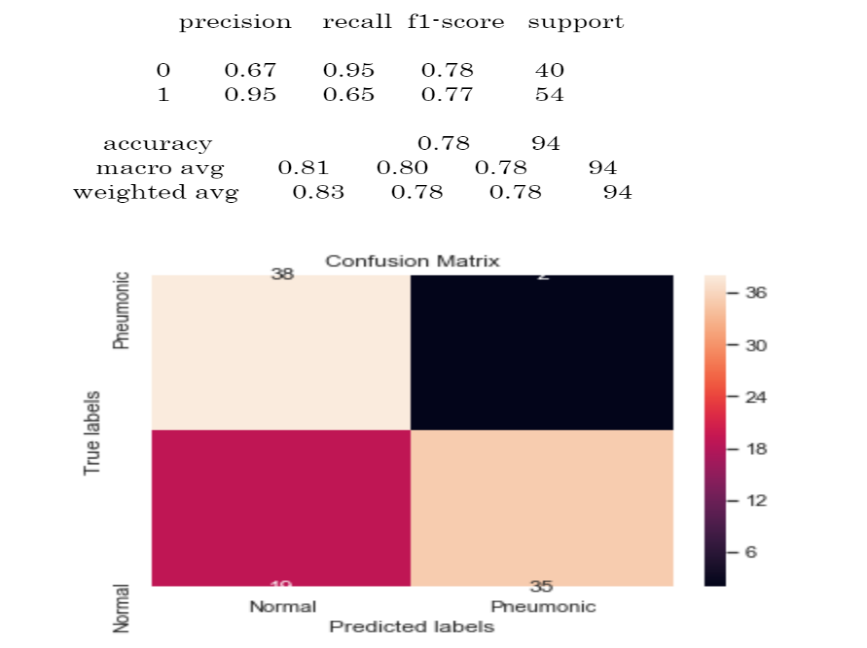


5.2. Performance for K-means on pneumonia using traditional method (PCA)



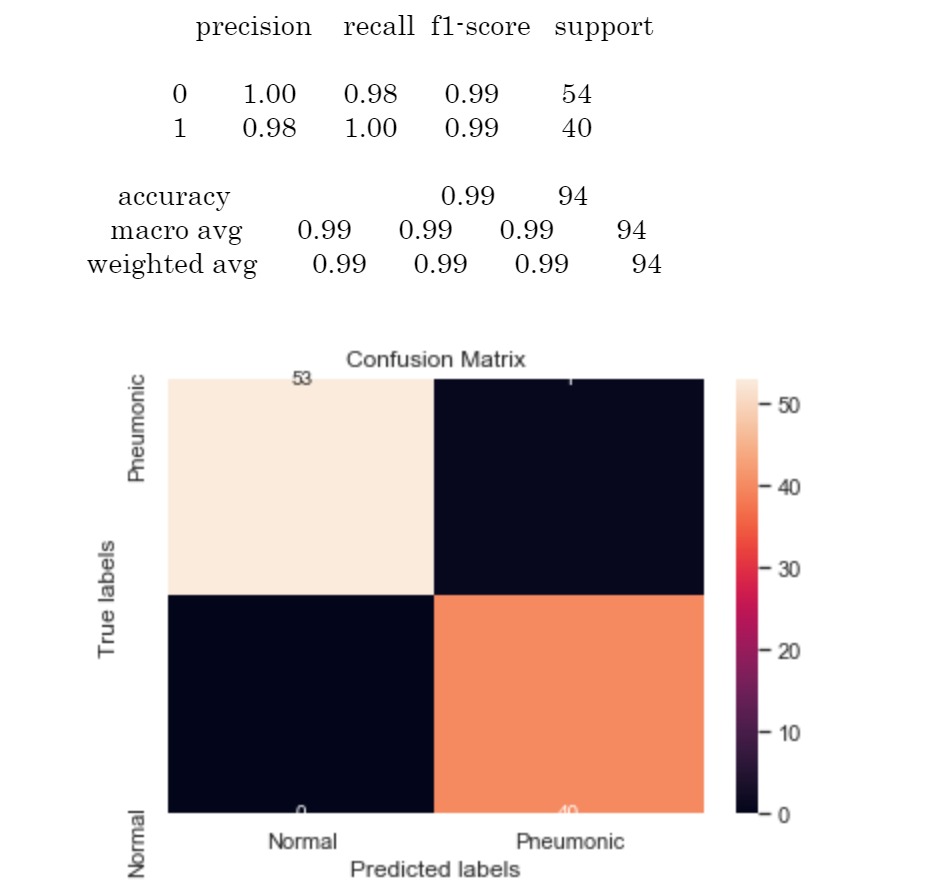
5.3. Distribution of images into clusters w. r. t. PC components

For the image matching technique (SIFT), considering all the features (pixel array) K-means gave the best result of:



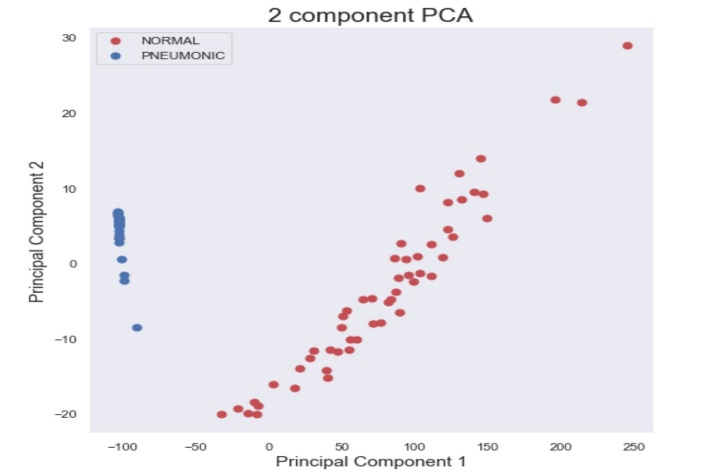
5.4. Performance for K-means on pneumonia using SIFT method

The final approach, transfer learning inception model (neural networks) gave the best results among all the 3 techniques and algorithms BIRCH, single linkage and ward gave the same result for all the performance metrics which are as shown below:



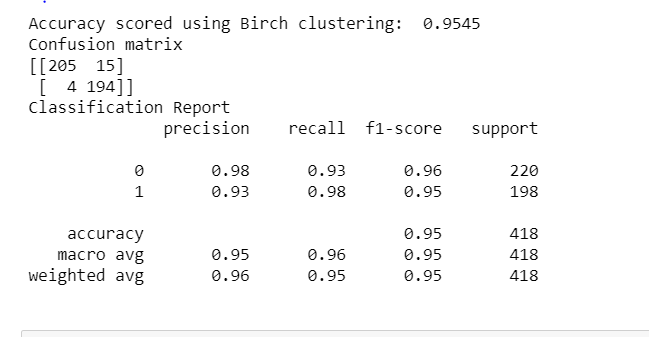
5.5. Performance on pneumonia using transfer learning method

Next, using PCA, the first 2 PC components were able to explain 98% of the variance. Hence, PCA for this can be considered. Clustering with these components gave the same result as with the original dataset for the clustering algorithms. The Distribution is as below:



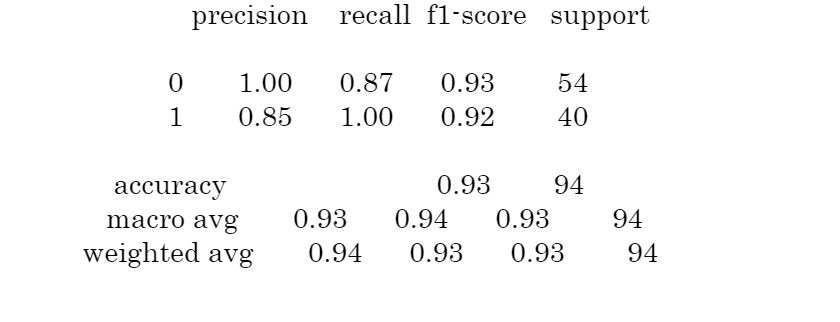
5.6. Distribution of images into clusters w. r. t. PC components

For **COVID** analysis, even the tradition method gave a very good result with BIRCH which is:



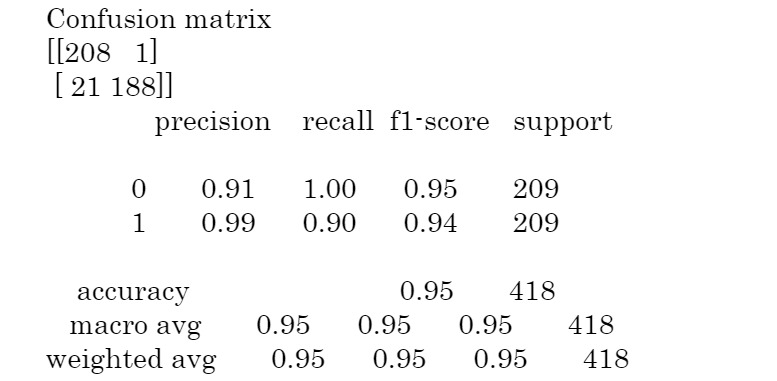
5.7. Performance on COVID dataset using traditional method

BIRCH gave the best result among the clustering algorithms for the SIFT technique also. The performance metrics are:



5.8. Performance on COVID dataset using traditional method

Transfer learning with inception model on COVID gave a very good result for almost all the models. Among them the outstanding one was BIRCH which are:



5.9. Performance on COVID dataset using transfer learning method

**5.1: Discussion of Results**

Inception technique gave the best result for the Pneumonia dataset. However, all the techniques used were able to give good results. For COVID dataset all the techniques gave good results (all the performance measures were >90%) using BIRCH algorithm. May be because the dataset is small even the traditional method gave the same result as a neural network method.

One limitation is that for COVID identification the dataset is very small (<300). Had there been a larger dataset, we would have known the reliability of the results. Also, other image matching techniques could be used. With respect to deep learning other transfer learning models or own models can be built.

**6: References**

1. <https://jakevdp.github.io/PythonDataScienceHandbook/05.11-k-means.html>
2. <https://towardsdatascience.com/prototyping-a-cnn-to-diagnose-covid-19-from-chest-x-ray-e1224a8e7795>
3. <https://www.kaggle.com/therealcyberlord/pneumonia-detection-using-cnn>
4. <https://www.researchgate.net/publication/228943913_Clustering_visually_similar_images_to_improve_image_search_engines#pf4>
5. <https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_feature2d/py_sift_intro/py_sift_intro.html>
6. <https://arxiv.org/ftp/arxiv/papers/1710/1710.02726.pdf>
7. <https://www.freecodecamp.org/news/getting-started-with-tesseract-part-ii-f7f9a0899b3f/>
8. <https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>
9. <https://towardsdatascience.com/detecting-covid-19-induced-pneumonia-from-chest-x-rays-with-transfer-learning-an-implementation-311484e6afc1>
10. <https://medium.com/@franky07724_57962/using-keras-pre-trained-models-for-feature-extraction-in-image-clustering-a142c6cdf5b1>
11. <https://towardsdatascience.com/image-clustering-using-transfer-learning-df5862779571>
12. <https://medium.com/@ksusorokina/image-classification-with-convolutional-neural-networks-496815db12a8>