

Sudan University of Science and Technology



Engineering College

College of Graduate Studies and Scientific Research

Biomedical Engineering Department

Classification and Detection of Coronavirus in Lung Images using Random Forests Algorithm

**التصنيف و الكشف لفيروس كورونا في الصور الطبيه للرئه باستخدام
خوارزمية الغابات العشوائية**

**Thesis submitted in partial fulfillment of the requirements for the
award of degree of Master of Science in Biomedical Engineering**

Submitted By: Rawaa Amir Awad Amir

Supervised By: Dr. Mohammed Yagoub Esmail

August 2021

الايه

بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِیْمِ

قال تبارك وتعالى:

(وَيَسْأَلُونَكَ عَنِ الرُّوحِ قُلِ الرُّوحُ مِنْ أَمْرِ رَبِّي وَمَا أُوتِيتُمْ مِنَ الْعِلْمِ إِلَّا قَلِيلًا)

الاسراء(85)

Dedication

*I dedicate this research with much love and appreciation;
To the candles of my lives. My beloved mother who have
always been there for me.*

*To my father who have always been the brick walls on
whom me can learn and depend on forever.*

*To my brothers and sister for the support when things were
up and mostly when there were down.*

*To my friends, family, colleagues and teachers in the Past
and presents and to everyone that touch my heart.*

ACKNOWLEDGMENTS

Foremost, I thank the almighty God for providing all these sources to carry out this research work.

I would like to express our sincere gratitude to our mentor and advisor Dr. Mohammed Yagoub, for his non-stopping support and guidance throughout the process of bringing out this project, I thank him for his patience, motivation, enthusiasm and immense knowledge that he provided without hesitation. His guidance helped in all times, I couldn't have wished for a better person to oversee our graduation project.. And of course I would like to thank and greet with all respect all doctors, teachers and staff of the department of biomedical engineering at Sudan University of Science and Technology.

I would like to thank my parents and brothers if it were not for their patience and continuous tenderness I would not be at this point.

Abstract

Coronavirus 2019 (COVID-19), which emerged in Wuhan, China and affected the whole world, has cost the lives of thousands of people. Manual diagnosis is inefficient due to the rapid spread of this virus. For this reason, automatic COVID-19 detection studies are carried out with the support of Random forest algorithms.

A research datasets consists 794 CT image slices was used to validate our proposed method. In this thesis, Firstly The pre-process done using filter to remove speckle noise and enhance the image as general. Then alveoli and COVID-19 segmentation are performed to be extracted from abdominal CT image using clustering texture (K-mean clustering) method. Secondly, texture feature information provided by GLCM is expected to differentiate between normal and abnormal tissue. Finally, COVID-19 detection is done on the segmented lung image using RF classifier, all the mentioned algorithm used in this project are robust and accurate more than the human visual system.

The result of proposed system 97.25% accuracy in distinguishing between normal alveoli and COVID-19.

المستخلص

ظهر فايروس كورونا في ووهان بالصين واثّر علي العالم وتسبب في مقتل الالف الاشخاص. ان التشخيص اليدوي غير فعال بسبب الانتشار السريع لهذا الفيروس. لهذا السبب ، يتم إجراء دراسات للكشف التلقائي عن فايروس كورونا بدعم من خوارزميات الغابات العشوائية.

في هذا البحث تم استخدام 794 من الصوره الطبيه المقطعيه وتم استخدامها للتحقق من صحه الطريقه المقترحه. اولا العمليه الاوليّه باستخدام مرشح لإزالة ضوضاء البقع وتحسين الصورة بشكل عام. يتم اجراء تجزئه الحويصلات الهوائيه و فايروس كورونا لاستخراجها من صورة التصوير المقطعي المحوسب في البطن باستخدام طريقة النسيج العنقودي. ثانيا : من المتوقع أن تكون معلومات إستخلاص المميزات التي تقدمها(جي ال سي ام) قادرة على التمييز بين الأنسجة الطبيعية وغير الطبيعية . أخيرا يتم إكتشاف فايروس كورونا على صورة الرئه باستخدام مصنف الغابات العشوائية. جميع الخوارزميات المذكورة المستخدمة في هذا المشروع قوية ودقيقة أكثر من النظام البصري البشري .

حقق النظام المقترح 97.25% في التمييز بين فايروس كورونا والحويصلات الهوائيه السليمه.

Table of content

الايه.....	I
DEDICATION	II
ACKNOWLEDGMENTS	III
ABSTRACTIV
المستخلص	V
Table of Contents	VI
LIST OF FIGURESXI
LIST OF TABLE	XII
ABBREVIATION	XIV
CHAPTER ONE.....	1
INTRODUCTION.....	1
1.1 General background.....	1
1.2 Problem statement.....	3
1.3 Objectives.....	3
1.3.1 General Objective.....	3
1.3.2 Specific Objectives.....	3
1.4 Methodology.....	3
1.6 Thesis layout.....	4
CHAPTER TOW.....	5
LITREATURE REVIEW.....	5
2.1 Overview of the most relevant systems ad method in CAD literature.....	5
2.1.1 A Comparative Analysis of Image De-noising Problem.....	5
2.1.2 A classification and analysis of pulmonary nodules in CT images using random forest.....	6
2.1.3 Detection of Coronavirus Disease Based on Deep Features and Support Vector Machine	6

2.1.4 Development of a Machine-Learning System to Classify Lung CT Images into Normal/COVID-19 Class.....	6
2.1.5 Classification of COVID-19 in Chest CT Images using Convolutional Support Vector Machines.....	7
2.1.6 Classification COVID-19 using Deep Features Fusion and Ranking Technique.....	7
2.1.7 Automatic COVID-19 CT segmentation using U-Net integrated spatial and channel attention mechanism.....	8
2.1.8 Identifying COVID19 from Chest CT Images based on Deep Convolutional Neural Networks.....	8
2.1.9 COVID-Classifer: an automated machine learning model to assist in the diagnosis of COVID-19 infection in chest X-ray images.....	8
2.1.10 COVID-19 detection from lung CT-scan images using transfer learning approach.....	9
2.1.11 Classification of COVID 19 in Chest CT Images using Convolutional Neural Network.....	9
2.2 Comparison between papers.....	10
2.3 Summary.....	11
CHAPTER THREE.....	12
THEORETICAL BACKGROUND.....	12
3.1 Respiratory system.....	12
3.1.1 Anatomy of the lung.....	12
3.1.2 Respiratory Zone Structures.....	13
3.1.2.1 Air Flows through an Extensive Airway System That Filters, Warms, and Humidifies the Air.....	15
3.2 coronavirus disease 2019 (COVID-19).....	16
3.2.1 Human Coronavirus Types.....	17
3.2.1.1 Common human coronaviruses.....	17
3.2.1.2 Other human coronaviruses.....	17
3.2.2 COVID-19 and Respiratory System Disorders.....	18

3.2.3 Causes and risk factor.....	18
3.2.4 Diagnosis of COVID 19.....	20
3.2.4.1 Viral test.....	20
3.2.4.2 Imaging.....	21
3.2.4.3 Coding.....	22
3.2.5 Staging of COVID 19.....	22
3.2.5.1 COVID-19 Stage 1: Viral Entry and Replication (Asymptomatic).....	22
3.2.5.2 COVID-19 Stage2 Viral Dissemination	22
3.2.5.3 COVID-19 Stage3 Multi-system Inflammation (Severe).....	23
3.2.5.4 COVID-19 Stage 4: Endothelial Damage, Thrombosis, and Multi-organ Dysfunction (Critical).....	23
3.3 Computed tomography (CT).....	24
3.4 Digital image processing.....	24
3.4.1 Types of Image.....	25
3.4.1.1 Greyscale image.....	25
3.4.1.2 RGB Image.....	25
3.4.2 Basic functions of Digital Image Processing.....	25
3.4.2.1 Image enhancement.....	26
3.4.2.2 Noise addition.....	26
3.4.2.2.1 Speckle noise.....	26
3.4.2.3 Filters.....	26
3.4.2.3.1 The Wiener filter.....	26
3.4.3 Image segmentation.....	28
3.4.3.1 Clustering-Based Segmentation Algorithms.....	28
3.4.3.1.1 K-means Clustering.....	28
3.4.4 Feature extraction.....	28
3.4.4.1 Texture feature.....	29

3.4.4.1.1 Contrast.....	30
3.4.4.1.2 Homogeneity.....	30
3.4.4.1.3 Entropy.....	31
3.4.4.1.4 Correlation.....	31
3.4.4.1.5 Autocorrelation.....	31
3.4.4.1.6 Energy.....	31
3.4.4.1.7 Variance.....	31
3.4.4.1.8 Sum average.....	31
3.4.4.1.9 Sum of variance.....	31
3.4.4.1.10 Sum of entropy.....	32
3.4.4.1.11 Difference entropy.....	32
3.4.4.1.12 Difference variance.....	32
3.4.4.1.13 Information measure of correlation 1.....	32
3.4.4.1 Equations of Texture feature.....	32
3.4.4 Classification.....	33
3.4.4.1 Random Forests.....	34
3.4.4.2 Random forest algorithm.....	34
3.4.4.2.1 Random forest creation.....	35
3.4.4.2.1 Random forest prediction.....	35
3.4.5 Background on Matlab and the Image Processing Toolbox.....	35
CHAPTER FOUR.....	37
Methodology.....	37
4.1 Introduction.....	37
4.1.1 Data collection.....	37
4.1.2 Image processing.....	38
4.1.3 Image segmentation.....	38
4.1.4 Feature extraction.....	39

4.1.5 Features Selection.....	40
4.1.6 Classification and Evaluation.....	40
4.1.6.1 Random Forest steps.....	40
4.1.7 Flow chart algorithm.....	42
CHAPTER FIVE.....	44
RESULT AND DISCUSSION.....	44
5.1 Result.....	44
5.1.1 Result of image preprocessing of non-COVID image.....	44
5.1.2 Result of preprocessing of COVID 19 image.....	45
5.1.3 Results of Segmentation.....	46
5.1.4 Results of feature extraction and selection.....	47
5.1.5 Result of classification.....	55
5.1.5.1 Performance measures.....	55
CHAPTER SIX.....	57
CONCLUSION AND RECOMMENDATION.....	57
6.1 Conclusion	57
6.2 Recommendation.....	57
Reference	58

LIST OF FIGURES

Figure (1.1): Block diagram of research methodology.....	4
Figure 3.1: structure of the lung.....	13
Figure 3.2: upper and lower respiratory tract.....	14
Figure 3.3: Human lung function.....	15
Figure 3.4: extensive branching.....	16
Figure 3.5: transmission of COVID 19.....	17
Figure 3.6: gas exchange and respiratory failure.....	19
Figure 3.7: COVID-19 diagnostic testing through real-time RT-PCR.....	21
Figure 3.8: A CT scan of a person with COVID-19.....	21
Figure 3.9: COVID-19 Staging.....	23
Figure (4.1): block diagram of research methodology.....	37
Figure 4.2: the GitHub website.....	38
Figure (4.2): Flow chart of algorithm.....	42
Figures (5.1) showing Image preprocessing of non-COVID.....	44
Figures (5.2) showing Image preprocessing of non-COVID 19.....	45
Figures (5.3) Shows the Segmentation stage.....	46
Figure (5.4) Image after diagnosis the red color explain the place of COVID 19... 	47
Figure 5.5 Autocorrelation feature.....	48
Figure (5.6) contrast feature.....	48
Figure (5.7) Correlation feature.....	48
Figure (5.8) Cluster Prominence feature.....	49
Figure (5.9) cluster shade feature.....	49
Figure (5.10) dissimilarity feature.....	49
Figure (5.10) Energy feature.....	50

Figure (5.12) Entropy feature.....	50
Figure (5.13) Homogeneity feature.....	51
Figure (5.14) Maximum probability feature.....	51
Figure (5.15) Variance feature.....	51
Figure (5.16) sum average.....	51
Figure (5.17) Sum variance feature.....	52
Figure (5.18) Sum entropy feature.....	52
Figure (5.19) Information measure of correlation1 feature.....	52
Figure (5.20) Information measure of correlation 2 feature.....	53
Figure (5.21) Inverse difference (INV) feature.....	53
Figure (5.22) Inverse difference normalized feature.....	53
Figure (5.23) solidity feature.....	54
Figure (5.24) Area feature.....	54
Figure (5.25): specifications of the random forest classifier.....	55
Figure 5.26: MATLAB performance.....	56

LIST OF TABLE

Table (2-1) the Comparison between papers10

Table (3.1): Equations of Texture feature.....33

ABBREVIATION

WHO	World Health Organization
COVID-19	coronavirus 2019
CT	computed tomography
RT-PCR	real time reverse transcription–polymerase chain reaction
CAD	Computer-aided detection
MATLAB	Matrix Laboratory
MSE	mean square error
PSNR	peak signal to noise ratio
RF	Random forest
SVM	support vector machine
GLCM	Gray Level Co-occurrence Matrix
ASM	Angular Second Moment
ANN	Artificial neural networks
INV	inverse difference
INN	Inverse difference normalized
NB	Naive Bayes
CNN	Convolutional Neural N

CHAPTER ONE

INTRODUCTION

1.1 General background

The World Health Organization (WHO) declared the new coronavirus called the COVID-19, which emerged in Wuhan, China and affected the whole world, pandemic, and it has brought the entire globe into a compulsory lockdown. Coronavirus is a family of RNA viruses that is capable of causing significant viral pathogens in humans and animals. Corona is medium-sized viruses with the largest viral RNA genome known also it can spread the virus to a human. As of two of March 2021, there have been more than 115 million confirmed cases of coronavirus worldwide, with about 2,550,726 of such cases resulting in the death of the infected patient. This is spread around 216 countries, areas, or territories. However, around 90 million infected patients have recovered worldwide [1].

Testing COVID-19 involves analyzing samples that indicate the present or past presence of severe acute respiratory syndrome-associated coronavirus. The test is done to detect the presence either of the virus or of antibodies produced in response to infection. COVID-19 diagnostic approach is mainly divided into two broad categories, a laboratory-based approach, which includes point of care testing, nucleic acid testing, antigens tests, and serology (antibody) tests. The other approach is using medical imaging diagnostic tools such as X-ray and computed tomography (CT) [2].

The laboratory-based tests are performed on samples obtained via nasopharyngeal swab, throat swabs, sputum, and deep airway material. The most common diagnostic approach is the nasopharyngeal swab the process is fast and is employed at the point of care. The nucleic acid test has low sensitivity between 60-71% [3]. Thoracic CT scan is the imaging modality of choice that plays a vital role in the management of COVID-19. It showed that radiologic methods could provide higher sensitivity than that of lab tests

CT scan involves transmitting X-rays through the patient's chest, which are then detected by radiation detectors and reconstructed into high-resolution medical images [4].

COVID-19 is a lung disease caused by a novel coronavirus first detected in late 2019. Its symptoms can range from mild to severe. Anyone can get COVID-19, but some individuals are more at risk for severe disease than others. The majority of people recover from COVID-19 within a few weeks, but it can be life threatening. Currently, three COVID-19 vaccines are authorized for emergency use. The best way to prevent illness is to avoid being exposed to the virus [5].

CT images show image features of multiple small patches and interstitial changes. It develops multiple ground glass shadows and infiltration shadows of the lungs. In severe cases, lung consolidation may occur, and pleural effusions are rare [6].

Generally Comparing to RT-PCR test, chest CT is relatively easy to operate and can provide fast diagnosis; moreover, chest CT has high sensitivity for screening COVID-19 infection and Fang *et al.*[7] compared the sensitivity of chest CT detection with nucleic acid detection by RT-PCR. 51 patients received initial and repeated RT-PCR tests. The standard is the diagnosis of COVID-19 infection finally confirmed by serial RT-PCR testing. Also have reported a lack of sensitivity in the initial RT-PCR test [8].

Studied 121 cases of chest CT studies obtained in the early, middle, and late infections of four centers in China. Studies have shown that the appearance of frosted glass on both sides and surrounding lungs is characteristic of the disease. Computer-aided diagnosis system uses imaging, medical image processing technology, and other means combined with computer analysis and calculation to assist in diagnosis. Many applications have been proposed in medical imaging, including segmentation and characterization tasks, also have reported a lack of sensitivity in the initial RT-PCR test [9].

This computer-aided diagnosis system uses medical image processing technology, and other means combined with computer analysis and calculation to assist in diagnosis. Many applications have been proposed in medical imaging, including preprocessing,

segmentation, feature extraction, and classification task use program MATLAB. The name MATLAB stands for Matrix Laboratory. MATLAB was written originally to provide easy access to matrix software developed by linear system package. MATLAB is a high-performance language for technical computing it is used this software program in create code to classify the COVID-19.

1.2 Problem statement

The radiologist to diagnosis COVID 19 is difficult to detect undesired cells visually from image directly also the restricted accessibility of this testing unit, the test of COVID 19 requires lots of time and effort the availability of testing kits pose another serious problem in terms of efficient detection of the disease. The human observer could make mistakes in taking decision and classifying is the key driver of designing CAD systems.

1.3 Objectives

The objectives of this research are general objective and specific objectives:

1.3.1 General Objective

The main purpose of this research is to design COVID- 19 classification system in lunge images to provide results that are more accurate in a fast and easy manner. Hence, this study aims to perform CAD system to give support to medical doctors in decision-making.

1.3.2 Specific Objectives are to:

- Design computer system for (processing, feature extraction and classification of COVID-19) to reduce the time of the radiologist in evaluation and classification.
- Classify COVID 19 into normal and abnormal cases.
- Improve the accuracy of COVID 19 classification task.
- Compare the performance and accuracy of the random forest with other machine-learning algorithm.

1.4 Methodology

This research consists of five phases as shown in the block diagram below. Selected data were obtained from the website and the hospital, this images passed through process like preprocessing include contrast stretching and wiener filter, gray level thresholding based k-mean clustering for segmentation, in feature extraction used twenty tow Haralik feature and finally, The RF attribute is applied to feature sets to find the most important features. These features were selected for use in the random forest classification and evaluation process.

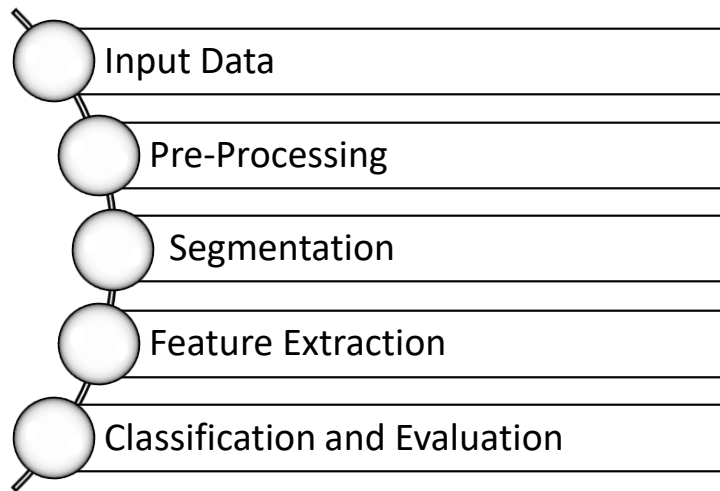


Figure (1.1): Block diagram of research methodology

1.6 Thesis layout

This research consists of five chapters: Chapter one is an introduction, chapter two deals of theoretical background and discusses the related literature review. The design and implementation of the classification system was explained in chapter three. The result and discussion were illustrated in chapter four, finally conclusion and recommendations were presented in chapter five.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview of the most relevant systems and methods in CAD literature

2.1.1 A Comparative Analysis of Image De-noising Problem

In 2020, Subrato *et al.* This paper mainly focuses on Gaussian noise, salt and pepper noise, uniform noise, speckle noise. Different filtering techniques can be adapted for noise reduction to improve the visual quality as well as a reorganization of images. Here four types of noises have been undertaken and applied to process images. CT or X-ray images must be noise free at the time of detecting COVID-19 from the input images with the help of various deep learning algorithms. Besides linear and nonlinear filtering methods like Gaussian filter, median filter, mean filter and Wiener filter applied for noise reduction as well as estimate the performance of filter through the parameters like mean square error (MSE), peak signal to noise ratio (PSNR), average difference value (AD) and maximum difference value (MD) to diminish the noises without corrupting the medical image data [10].

2.1.2 A classification and analysis of pulmonary nodules in CT images using random forest

In 2018, S.santhosh baboo and E.iyyapparaj. proposed a novel Computer-aided detection (CAD) system based on a Contextual clustering combined with region growing for assisting radiologists in early identification of lung cancer from computed tomography(CT) scans. Used wiener and morphological filter with conventional thresholding approach; this proposed work uses Contextual Clustering, which yields a more accurate segmentation of the lungs from the chest volume. Following segmentation GLCM and LBP features are extracted which are then classified using three different

classifiers namely Random forest, SVM and k-NN. From performance metrics obtained it is found that Random Forest based classifier outperforms other classifiers [11].

2.1.3 Detection of Coronavirus Disease Based on Deep Features and Support Vector Machine

In 2020, Prabira kumar Sethy *et al.* In this paper, the deep feature plus support vector machine (SVM) based methodology is suggested for detection of coronavirus infected patient using X-ray images. For classification, SVM is used instead of deep learning based classifier. The deep features from the fully connected layer of CNN model are extracted and fed to SVM for classification purpose. The SVM classifies the corona affected X-ray images from others. SVM is evaluated for detection of COVID-19 using the deep features of different 13 number of CNN models. The SVM produced the best results using the deep feature of ResNet50. The classification model, i.e. ResNet50 plus SVM achieved accuracy, sensitivity, FPR and F1 score of 95.33%,95.33%,2.33% and 95.34% respectively for detection of COVID-19. Again, the highest accuracy achieved by ResNet50 plus SVM is 98.66%. As the data set is in hundreds, the classification based on SVM is more robust compared to the transfer learning approach. In traditional image classification method, LBP plus SVM achieved 93.4% of accuracy [12].

2.1.4 Development of a Machine-Learning System to Classify Lung CT Images into Normal/COVID-19 Class

In 2020, Seifedine et al. presented Machine-Learning-System (MLS) to detect the COVID-19 infection using the CT scan Slices (CTS). This MLS implements a sequence of methods, such as multi-thresholding, image separation using threshold filter, feature-extraction, feature-selection, feature-fusion and classification. The threshold filter separates the image into two segments based on a chosen threshold. The texture features of these images are extracted, refined and selected using the chosen procedures. Finally, a two-class classifier system is implemented to categorize the chosen into normal/COVID-19 group. In this work, the classifiers, such as Naive Bayes (NB), k-

Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF) and Support Vector Machine with linear kernel (SVM) are implemented and the classification task is performed using various feature vectors. The experimental outcome of the SVM with Fused-Feature-Vector (FFV) helped to attain a detection accuracy of 89.80% [13].

2.1.5 Classification of COVID-19 in Chest CT Images using Convolutional Support Vector Machines

In 2020, Umut Özkaya *et al.* Presented a deep learning model that detects COVID-19 cases with high performance. The proposed method is defined as Convolutional Support Vector Machine (CSVM) and can automatically classify Computed Tomography (CT) images. Unlike the pre-trained Convolutional Neural Networks (CNN) trained with the transfer learning method, the CSVM model is trained as a scratch. To evaluate the performance of the CSVM method, the dataset is divided into two parts as training (75%) and testing (25%). The CSVM model consists of blocks containing three different numbers of SVM kernels. Results: When the performance of pre-trained CNN networks and CSVM models is assessed, CSVM model shows the highest performance with 94.03% ACC, 96.09% SEN, 92.01% SPE, 92.19% PRE, 94.10% F1-Score, 88.15% MCC and 88.07% Kappa metric values. The proposed method is more effective than other methods. It has proven in experiments performed to be an inspiration for combating COVID and for future studies [14].

2.1.6 Classification COVID-19 using Deep Features Fusion and Ranking Technique

In 2020, Umut Ozkaya, Saban Ozturk and Mucahid Barstugan. Proposed as fusing and ranking deep features to detect COVID-19 in early phase. 16x16 (Subset-1) and 32x32 (Subset-2) patches were obtained from 150 CT images to generate sub-datasets. Within the scope of the proposed method, 3000 patch images have been labelled as COVID-19 and No finding for using in training and testing phase. Feature fusion and ranking method have been applied in order to increase the performance of the proposed method. Then, the processed data was classified with a Support Vector Machine (SVM). According to other pre-trained Convolutional Neural Network (CNN)

models used in transfer learning, the proposed method shows high performance on Subset-2 with 98.27% accuracy, 98.93% sensitivity, 97.60% specificity, 97.63% precision, 98.28% F1-score and 96.54% Matthews Correlation Coefficient (MCC) metrics [15].

2.1.7 Automatic COVID-19 CT segmentation using U-Net integrated spatial and channel attention mechanism

In 2020, Tongxue Zhou, Stéphane Canu, and Su Ruan. proposed a U-Net based segmentation network using attention mechanism. As not all the features extracted from the encoders are useful for segmentation, they propose to incorporate an attention mechanism including a spatial attention module and a channel attention module, to a U-Net architecture to re-weight the feature representation spatially and channel-wise to capture rich contextual relationships for better feature representation. In addition, the focal Tversky loss is introduced to deal with small lesion segmentation. The experiment results, evaluated on a COVID-19 CT segmentation dataset where 473 CT slices are available, demonstrate the proposed method can achieve an accurate and rapid segmentation result on COVID-19. The method takes only 0.29 second to segment a single CT slice. The obtained Dice Score and Hausdorff Distance are 83.1% and 18.8, respectively [16].

2.1.8 Identifying COVID19 from Chest CT Images based on Deep Convolutional Neural Networks

In 2020, Arnab Kumar Mishra et al. proposed various Deep CNN based approaches are explored for detecting the presence of COVID19 from chest CT images. A decision fusion based approach is also proposed, which combines predictions from multiple individual models, to produce a final prediction. Experimental results show that the proposed decision fusion based approach is able to achieve above 86% results across all the performance metrics under consideration, with average AUROC and F1-Score being 0.883 and 0.867, respectively. The experimental observations suggest the potential applicability of such Deep CNN based approach in real diagnostic scenarios, which could be of very high utility in terms of achieving fast testing for COVID19 [17].

2.1.9 COVID-Classifier: an automated machine learning model to assist in the diagnosis of COVID-19 infection in chest X-ray images

In 2021, Abolfazl et al. proposed they have trained several deep convolutional networks with introduced training techniques for classifying X-ray images into three classes: normal, pneumonia, and COVID-19, based on two open-source datasets. Our data contains 180 X-ray images that belong to persons infected with COVID-19, and they attempted to apply methods to achieve the best possible results. In this research, we introduce some training techniques that help the network learn better, when they have an unbalanced dataset. They also propose a neural network that is a concatenation of the Xception and ResNet50V2 networks. This network achieved the best accuracy by utilizing multiple features extracted by two robust networks. For evaluating our network, they have tested it on images to report the actual accuracy achievable in real circumstances. The average accuracy of the proposed network for detecting COVID-19 cases is 99.50%, and the overall average accuracy for all classes is 91.4% [18].

2.1.10 COVID-19 detection from lung CT-scan images using transfer learning approach

In 2021, Arpita Halder, and Bimal Datta. Described the development of a DL framework that includes pre-trained models (DenseNet201, VGG16, ResNet50V2, and MobileNet) as its backbone, known as KarNet. To extensively test and analyze the framework, each model was trained on original and manipulated datasets. Among the four pre-trained models of KarNet, the one that used DenseNet201 demonstrated excellent diagnostic ability, with AUC scores of 1.00 and 0.99 for models trained on un augmented and augmented data sets, respectively. Even after considerable distortion of the images DenseNet201 achieved an accuracy of 97% for the test dataset, MobileNet, and VGG16 ,which achieved accuracies of 96%, 95%, and 94%, respectively [19].

2.11 Classification of COVID 19 in Chest CT Images using Convolutional Neural Network

In 2021, Manikandan, Jabin Alf, Sherin, Aadhithya, and Senthil Kumar. Presented The test for the COVID-19 involves analysing the throat swab sample which may take days if not a week and by the time the results come the infection would have spread. Hence there is a need to improve the testing procedure for COVID-19. In this paper, they have come up with an automated Image Analysis technique to diagnose COVID-19 using the chest Computed Tomography images of the chest that uses a Convolutional Neural Network. The adopted method is also shown to be more effective as compared to conventional techniques of analyzing images manually. The issue of high time requirement is compensated in this technique and this method has proven to be quite competent having been able to achieve an accuracy as high as 98%. This proposed method is simple and cost-effective to be implemented both in rural and urban areas [20].

2.2 Comparison between papers

Table (2-1) the Comparison between papers

Reference	Implemented investigative procedure	Findings
Dr. S.SANTHOSH BABOO and E.IYYAPPARAJ	This paper classification of pulmonary nodules in CT images using three different classifiers namely Random forest, SVM and k-NN.	Provided a classification Random Forest based classifier outperforms other classifiers.
Wu et al.	This paper implemented a deep-learning procedure for the segmentation and classification of COVID-19 infection from CT images attained from 200 patients	Provided the dice score of 78.3% for segmentation. Further, helped to achieve an average sensitivity of 95.0% and a specificity of 93.0% during the classification.
Rahimzadeh and Attar	Implements a modified deep convolutional neural network for the COVID-19 diagnosis. (chest x ray)	This work provided a classification accuracy of 99.56% for the disease class and average accuracy of 91.4%.

Ozkaya et al.	This paper implemented a deep learning based on features fusion and ranking technique.	Helped to attain better values of accuracy (98.27%), sensitivity (98.93%), specificity (97.60%), precision (97.63%), and F1-score (98.28%).
---------------	--	---

2.3 Summary

From many scientific papers and literature studies between (2018_2021). all papers related to our topic seem to be following the same general steps with modification along the way. It was noticed that according to the dataset, processing is done to enhance the images and remove noise. As for segmentation, K-means clustering was mostly used because it is less complicated. Most papers extracted statistical features and some other additional features. The AAN and machine learning (SVM), and random forest classifier were mostly used in previous studies but the random forest is not use in classification of COVID 19 only used in pulmonary nodules, it was noticed that RF gave high accuracies even with small dataset.

CHAPTER THREE

THEORETICAL BACKGROUND

3.1 Respiratory system

Respiratory system is the network of organs and tissues that help you breathe. This system helps your body absorb oxygen from the air so your organs can work. It also cleans waste gases, such as carbon dioxide, from your blood. Common problems include allergies, diseases or infections. The human respiratory tract can be divided into 24 generations. The lung parenchyma distal to the terminal bronchiole is known as the acinus, and it constitutes the functional unit of the lung where gas exchange occurs. Regarding physiological functions, the contiguous airway from the trachea to the terminal bronchioles is called the conducting zone, and the area from the respiratory bronchioles to the alveolar sacs is called the transitional and respiratory zone. The human respiratory system could also be said to include the nasal cavity, associated sinuses, nasopharynx, oral cavity, oropharynx, and larynx [21-23].

3.1.1 Anatomy of the lung

The lungs are the major organs of the respiratory system, and are divided into sections, or lobes. The right lung has three lobes and is slightly larger than the left lung,

which has two lobes. The lungs are separated by the mediastinum. This area contains the heart, trachea, esophagus, and many lymph nodes. The lungs are covered by a protective membrane known as the pleura and are separated from the abdominal cavity by the muscular diaphragm. Each lung has an oblique fissure separating the upper lobes from the lower lobes and the right lung has a horizontal fissure that separates the right upper lobe from the middle lobe. With each inhalation, air is pulled through the windpipe (trachea) and the branching passageways of the lungs (the bronchi), filling thousands of tiny air sacs (alveoli) at the ends of the bronchi [21] .

These sacs, which resemble bunches of grapes, are surrounded by small blood vessels (capillaries). Oxygen passes through the thin membranes of the alveoli and into the bloodstream. The red blood cells pick up the oxygen and carry it to the body's organs and tissues. As the blood cells release the oxygen, they pick up carbon dioxide, a waste product of metabolism. The carbon dioxide is then carried back to the lungs and released into the alveoli. With each exhalation, carbon dioxide is expelled from the bronchi out through the trachea [21].

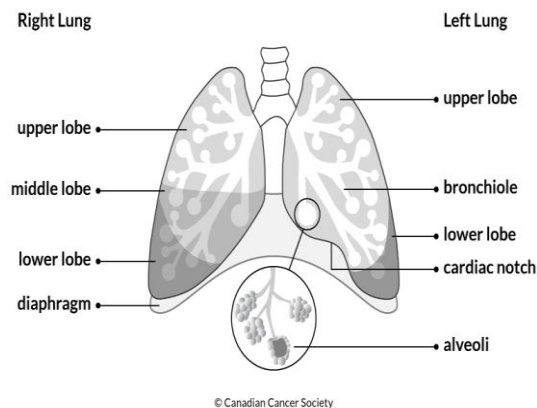


Figure 3.1: structure of the lung [21].

3.1.2 Respiratory Zone Structures

The respiratory zones of the airways include the respiratory bronchioles and alveoli. The airway wall in the respiratory zones of the airways is much thinner,

therefore maximizing gaseous exchange between the oxygenated inspired air and the gas dissolved in the pulmonary capillaries. Unlike the epithelial cells lining the conducting zones of the airways, the epithelial cells in the respiratory zones of the airways have reduced height and, as one descends into the respiratory zones, the epithelial cells are mostly composed of cuboidal and non-ciliated cells [24].

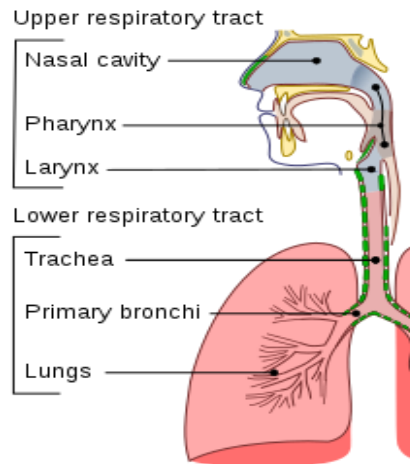


Figure 3.2: upper and lower respiratory tract [24].

A study of the obstructive spirometry patterns in the lungs of 6167 adult patients from northern India showed considerable variability between patients' forced expiratory volume in 1 second (FEV1), expressed as a percentage of predicted, and peak expiratory flow (PEF), expressed as a percentage of predicted. Analysis also showed that PEF expressed in this way overestimated the FEV1 percentage in patients with less severe obstruction and underestimated the FEV1 percentage in those with more severe obstruction, and inhalation rate differed in many patients with airflow limitation. Therefore, it is unjustified to assume general parity between PEF and FEV1 expressed as percentages and these terms should not be used interchangeably [22].

Given that respiratory blood-gas disorders primarily result from alterations of CO₂ levels, consideration must be also to alterations in O₂ levels as patients with respiratory acid-base disorders may also become hypoxemic. Oxygen transport is initiated by contraction of the diaphragm with consequent movement of inspired gas down the

continually branching airways until the transitional and respiratory bronchioles, alveolar ducts, and alveoli are reached. Within this respiratory zone, alveolar ventilation and gas exchange occur as oxygen moves down its concentration gradient and into the red blood cells. The partial pressure of oxygen in the red blood cells approximates that of alveolar gas within the first third of the lung capillaries, primarily due to the lung's considerable diffusion capabilities [23].

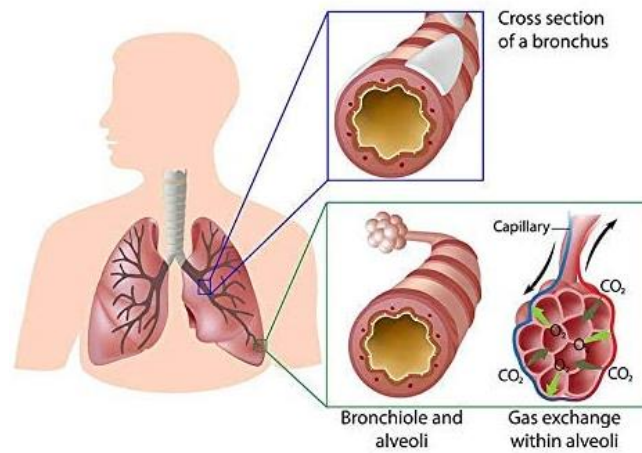


Figure 3.3: Human lung function [23].

3.1.2.1 Air Flows through an Extensive Airway System That Filters, Warms, and Humidifies the Air

Inspired air enters through the nose or mouth, or both, where it is conditioned: it is filtered, warmed, and humidified. Many hairs and sticky mucus trap large dust particles and thus filter the inspired air. Outcroppings of tissue in the nasal cavity called turbinates expose the air to a large surface area and mix the air within the nasal passages, humidifying and warming the air to 37°C.

After leaving the nasal passages, air travels through the throat, or pharynx, and then through the larynx or voice box. It then enters the trachea, which subsequently branches many times. The pharynx, larynx, and early generations of the airways leading to the

lungs do not participate in gas exchange. The movement of the mucus is like a “mucus escalator” that constantly cleans the lungs. The mucus brought to the mouth is typically expectorated (spat out) or swallowed. Persons with cystic fibrosis produce thick mucus that cannot be easily removed. Part of their treatment is a vigorous thumping of the chest to clear out the lungs [25].

The airways beyond the trachea constitute the tracheobronchial tree. Each branching of the airways produces the next generation, like a family tree. The first few generations conduct the air toward the gas exchanging regions, but they do not themselves exchange gases. The larger branches are the bronchi and bronchioles. These conductive airways become progressively smaller and more numerous as they branch, leading eventually to terminal bronchioles, respiratory bronchioles, and alveolar ducts, ending in dead-end sacs called alveoli (Figure 3.3). The alveoli are small sacs, about 0.3 mm in diameter, where tiny distances separate the blood from the alveolar gases. The lungs contain some 300×10^6 alveoli. Thus, the airways from nose to alveoli consist of a conducting zone of generations 0–16 and a respiratory zone in series with the conducting zone, with the respiratory zone consisting of generations 17–23 [26].

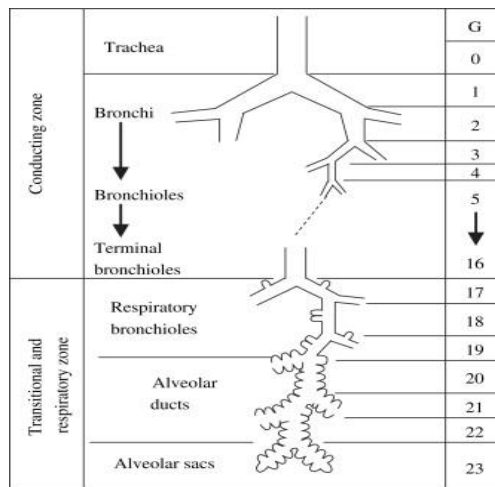


Figure 3.4: extensive branching [25].

3.2 coronavirus disease 2019 (COVID-19)

The severe acute respiratory syndrome coronavirus-2 a serious human pathogen in late 2019, causing the disease coronavirus disease 2019 (COVID-19). The World Health Organization designated coronavirus disease 2019 (COVID-19) as the name of the human disease caused by severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2). The most common clinical presentation of severe COVID-19 is acute respiratory failure consistent with the acute respiratory distress syndrome respiratory syndrome coronavirus-2 emerged as a serious human pathogen in late 2019, causing the disease coronavirus disease 2019 (COVID-19). The most common clinical presentation of severe COVID-19 is acute respiratory failure consistent with the acute respiratory distress syndrome. Airway, lung parenchymal, pulmonary vascular and respiratory neuromuscular disorders all feature in COVID-19 [27].



Figure 3.5: transmission of COVID 19 [27].

3.2.1 Human Coronavirus Types

Coronaviruses are named for the crown-like spikes on their surface. There are four main sub-groupings of coronaviruses, known as alpha, beta, gamma, and delta. Human coronaviruses were first identified in the mid-1960s. The seven coronaviruses that can infect people are:

3.2.1.1 Common human coronaviruses

- 1) 229E (alpha coronavirus)
- 2) NL63 (alpha coronavirus)
- 3) OC43 (beta coronavirus)
- 4) HKU1 (beta coronavirus)

3.2.1.2 Other human coronaviruses

- 1) MERS-CoV (the beta coronavirus that causes Middle East Respiratory Syndrome, or MERS)
- 2) SARS-CoV (the beta coronavirus that causes severe acute respiratory syndrome, or SARS)
- 3) SARS-CoV-2 (the novel coronavirus that causes coronavirus disease 2019, or COVID-19)

People around the world commonly get infected with human coronaviruses 229E, NL63, OC43, and HKU1. Sometimes coronaviruses that infect animals can evolve and make people sick and become a new human coronavirus. Three recent examples of this are 2019-nCoV, SARS-CoV, and MERS-CoV [28].

The respiratory tract epithelium is the key entry point for beta-coronaviridae, which includes SARS-CoV-2, MERS-CoV (Middle East respiratory syndrome-related coronavirus), and SARS-CoV, into the human host. The airway epithelium acts as a barrier to pathogens and particles, preventing infection and tissue injury by the secretion of mucus and the action of mucociliary clearance while maintaining efficient airflow.

In vitro data from SARS-CoV indicate that the ciliated airway epithelium serves as a primary site for viral infection; however, whether these airway epithelial cells express sufficient ACE2 to permit viral entry is controversial [29].

3.2.2 COVID-19 and Respiratory System Disorders

Viral pneumonia is the most frequent serious clinical manifestation of COVID-19, prominently featuring fever, cough, dyspnea, hypoxemia, and bilateral infiltrates on chest radiography. Dry cough is more common than a productive cough. Dyspnea appears after a median time of 5 to 8 days. Severe hypoxemic respiratory failure consistent with the Berlin definition of the acute respiratory distress syndrome (ARDS) occurs in a significant proportion of patients with COVID-19 pneumonia. Patients who require mechanical ventilation have a high risk of death [30].

3.2.3 Causes and risk factor

COVID-19 affects all components of the respiratory system, including the neuromuscular breathing apparatus, the conducting airways, the respiratory airways and alveoli, the pulmonary vascular endothelium, and pulmonary blood flow. The presence of viral particles in the nasal epithelium is the underlying rationale for obtaining nasopharyngeal material for polymerase chain reaction–based detection of the SARS-CoV-2 genome. Current polymerase chain reaction–based diagnostic tests for SARS-CoV-2 infection lack quantification of viral load and have variable negative and positive predictive value [31].

After entering and replicating within the nasal mucosa, SARS-CoV-2 travels to the conducting airways, where it triggers an immune and inflammatory response, manifesting in clinical signs and symptoms of COVID-19. Although SARS-CoV-2 infection often begins in the upper airway epithelium, in a subset of patients, the virus infects or injures the alveolar epithelium diffusely, resulting in markedly impaired gas exchange and respiratory failure [32].

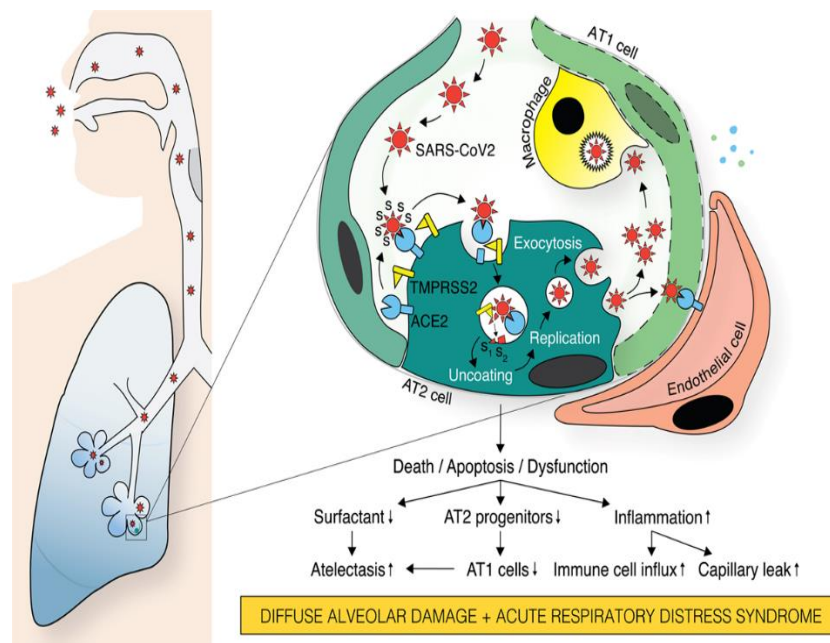


Figure 3.6: gas exchange and respiratory failure [32].

People who are older have a higher risk of serious illness from COVID-19, and the risk increases with age. People who have existing medical conditions also may have a higher risk of serious illness. Certain medical conditions that may increase the risk of serious illness from COVID-19 include serious heart diseases, such as heart failure, coronary artery disease or cardiomyopathy, Cancer, Chronic obstructive pulmonary disease (COPD), Type 1 or type 2 diabetes ,Overweight, obesity or severe obesity ,High blood pressure, Smoking, Chronic kidney disease, Sickle cell disease or thalassemia, Weakened immune system from solid organ transplants, Pregnancy, Asthma, Chronic lung diseases such as cystic fibrosis or pulmonary fibrosis, Liver disease, and Dementia. This list is not all-inclusive. Other underlying medical conditions may increase your risk of serious illness from COVID-19 [33].

The virus that causes COVID-19 spreads easily among people, and more continues to be discovered over time about how it spreads. Data has shown that it spreads mainly from person to person among those in close contact (within about 6 feet, or 2 meters). The virus spreads by respiratory droplets In some situations, the COVID-19 virus can spread by a person being exposed to small droplets or aerosols that stay in the air for several minutes or hours — called airborne transmission. It's not yet known how common it is for the virus to spread this way. It can also spread if a person touches a surface or object with the virus on it and then touches his or her mouth, nose or eyes, but the risk is low[34].

3.2.4 Diagnosis of COVID 19

COVID-19 can provisionally be diagnosed based on symptoms and confirmed using reverse transcription polymerase chain reaction (RT-PCR) or other nucleic acid testing of infected secretions [35]. Along with laboratory testing, chest CT scans may be helpful to diagnose COVID-19 in individuals with a high clinical suspicion of infection. There are varieties of tests to diagnosis COVID-19 in lung:

3.2.4.1 Viral test

The standard methods of testing for presence of SARS-CoV-2 are nucleic acid tests. The test is typically done on respiratory samples obtained by a nasopharyngeal swab; however, a nasal swab or sputum sample may also be used. Results are generally available within hours. The WHO has published several testing protocols for the disease [36].

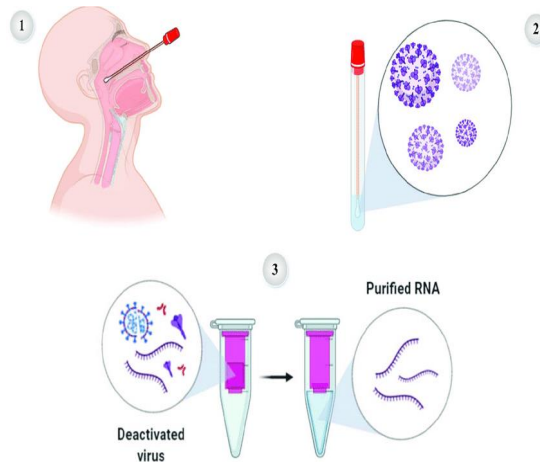


Figure 3.7: COVID-19 diagnostic testing through real-time RT-PCR [36].

3.2.4.2 Imaging

Chest CT scans may be helpful to diagnose COVID-19 in individuals with a high clinical suspicion of infection but are not recommended for routine screening. Characteristic imaging features on chest radiographs and computed tomography (CT) of people who are symptomatic include asymmetric peripheral ground-glass opacities without pleural effusions. A large study in China compared chest CT results to PCR and demonstrated that though imaging is less specific for the infection, it is faster and more sensitive [37].

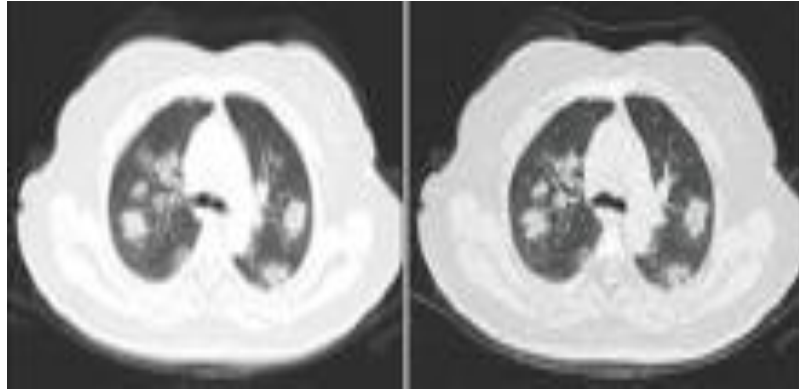


Figure 3.8: A CT scan of a person with COVID-19 [37].

3.2.4.3 Coding

The WHO assigned emergency ICD-10 disease codes U07.1 for deaths from lab-confirmed SARS-CoV-2 infection and U07.2 for deaths from clinically or epidemiologically diagnosed COVID-19 without lab-confirmed SARS-CoV-2 infection [38].

3.2.5 Staging of COVID 19

Disease staging is a method for measuring the progression and severity of an illness using objective clinical and molecular criteria. Staging provides valuable frameworks and benchmarks for clinical decision-making in patient management, improved prognostication, and evidence-based treatment selection. Through such a process, staging defines discrete points in the course of a particular disease that are clinically detectable and reflect present risk, potential long-term effects, and likelihood of death. Staging systems provide valuable frameworks and benchmarks for clinical decision-making in patient management, improved prognostication, and evidence-based treatment selection [39].

3.2.5.1 COVID-19 Stage 1: Viral Entry and Replication (Asymptomatic)

Asymptomatic disease has been well described since the earliest days of the SARS-CoV-2 pandemic. Children and younger adults are more likely to be in this group, though asymptomatic nucleic acid amplification test (NAAT)-positive cases have been seen in all age and risk groups to varying degrees. Treatment is not recommended for this group, though individuals may be eligible for temporal surveillance studies and/or clinical trials focused on prevention of symptoms, reduction of infectivity, and disease progression [40].

3.2.5.2 COVID-19 Stage 2: Viral Dissemination (Mild or Moderate)

The patients present one or more of the following symptoms: fever or chills, cough, shortness of breath, fatigue, muscle or body aches, headache, new loss of taste or smell, sore throat, congestion or runny nose, nausea or vomiting, and diarrhea. Stage 2 requires testing and confirmation of disease, isolation, and infection prevention precautions, as most people are likely highly infectious during their acute period of symptoms [41].

3.2.5.3 COVID-19 Stage 3: Multi-system Inflammation (Severe)

In this stage, often about one to two weeks after symptom onset, patients experience worsening dyspnea and hypoxia, along with subclinical elevations in indicators of organ damage. Patients may have reduced benefit from antibody-based treatments, and instead therapy should focus on providing oxygen support, anti-inflammatory and immunomodulatory therapies, anti-thrombotic, and clinical trials, including experimental therapies such as mesenchymal stem cells [42].

3.2.5.4 COVID-19 Stage 4: Endothelial Damage, Thrombosis, and Multi-organ Dysfunction (Critical)

Clinical phenomena include severe hypoxemic respiratory failure associated with multi-organ failure, including myocardial injury as evidenced by troponemia, with cardiac structural abnormalities and arrhythmias, severe acute kidney injury requiring

renal replacement therapy, acute neurological disease, venous and arterial thromboembolic events, and severe metabolic derangements, such as persistent hyperglycemia and ketosis. In this stage the underlying pathophysiology is primarily driven by the inflammatory response and coagulopathy rather than direct viral injury [43].

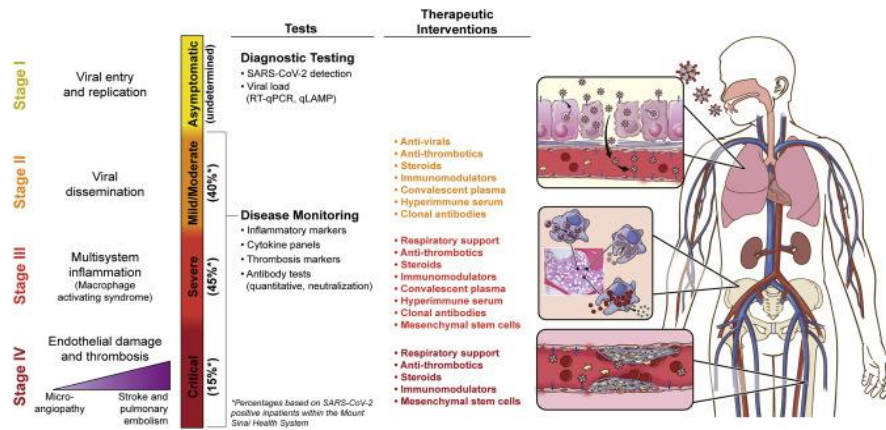


Figure 3.9: COVID-19 Staging.

3.3 Computed tomography (CT)

CT scans is a noninvasive medical examination or procedure that uses specialized X-ray equipment to produce cross-sectional images of the body. Each cross-sectional image represents a “slice” of the person being imaged, like the slices in a loaf of bread. These cross-sectional images are used for a variety of diagnostic and therapeutic purposes.

CT scans can be performed on every region of the body for a variety of reasons (e.g., diagnostic, treatment planning, interventional, or screening). Most CT scans are performed as outpatient procedures. This combining of the images allows greater soft tissue detail to be displayed. Each individual picture of a CT study is referred to as a section or an axial “slice.” This is because the picture must be interpreted as if the patient has been completely sectioned in an axial plane, like a loaf of bread, with the viewer looking at the section from the feet toward the head [44].

3.4 Digital image processing

The image processing is an advanced technology that enables you to manipulate digital images through computer software. It is the subfield of signal processing, which focuses primarily on images. Digital image processing allows the user to take the digital image as an input and perform the different algorithm on it to generate an output. These algorithms may vary from image to image according to the desired output image. Adobe Photoshop is the most popular software that uses digital image processing to edit or manipulate images.

One of the key benefits of integrating regularly is that you simply can detect errors quickly and locate them more easily. As each change introduced is usually small, pinpointing the precise change that introduced a defect is often done quickly. The three primary phases that constitute image processing are importing the image using picture acquisition tools, picture processing and handling, and output in which image or report based on image analysis may be changed.

3.4.1 Types of Image

An image is a 2D array of numeric values called pixels. These pixels carry a value that indicates the amount of light for that particular pixel. The pixel carry value represents information about the number of intensities present in the image. The value 0 represents that color black, the value 1 represents the color white. The images are further characterized in two types:

3.4.1.1 Greyscale image

The color grey lies within the white and black range. The images whose pixels carry values between zero to one are characterized as greyscale images.

3.4.1.2 RGB Image

RGB stands for Red, Green, and Blue. Any other color can be derived from these three primary colors. Each pixel of a colored image carries different 16 or 24-bit color

values. These 16 or 24 bits are further divided into three values that correspond to the RGB values. The combination of the RGB forms the exact colors of the pixel [45].

3.4.2 Basic functions of Digital Image Processing

Here are some basic functions that can be performed on an image, which will change the characteristic of the image.

3.4.2.1 Image enhancement

The image enhancement function uses its algorithms to improve image features. It adjusts the image in such a form that all the results are more suitable for further analysis. Also, this helps the user to extract hidden features by using techniques like sharpening. The function improves the visualization, removes the unwanted parts, deblurs the image, and much more.

3.4.2.2 Noise addition

Noise is the unwanted parts of the image. The noise is added to an image for testing purposes, it helps the user to test the efficiency of the noise removal filters. There are different types of noises, some of which are like Salt and pepper noise, Gaussian noise, Speckle noise, Motion blur.

3.4.2.2.1 Speckle noise

An active radar sensor gives off a burst of coherent radiation, which reflects from the target, unlike a passive microwave sensor, which simply receives the low-level radiation naturally emitted by targets. The waves emitted by the active sensors are in phase unless they strike a target. Upon interaction with the target, these waves get out of phase. These out of phase waves interact to produce a mix of light and dark pixel in an image. This disturbance is referred to as the Speckle Noise.

3.4.2.3 Filters

In digital image processing, filters are used to perform a different function on the image, such as removing noise, enhancing the image, detecting edges, and much more.

There are different types of noises, and they require different filters to remove them; some of them are Median filter, Laplacian filter, Gaussian filter, and wiener filter.

3.4.2.3.1 The Wiener filter

The Wiener filter is the MSE-optimal stationary linear filter for images degraded by additive noise and blurring. Calculation of the Wiener filter requires the assumption that the signal and noise processes are second-order stationary (in the random process sense). For this description, only noise processes with zero mean will be considered (this is without loss of generality). Wiener filters are usually applied in the frequency domain. Given a degraded image $x(n,m)$, one takes the Discrete Fourier Transform (DFT) to obtain $X(u,v)$. The original image spectrum is estimated by taking the product of $X(u,v)$ with the Wiener filter $G(u,v)$:

$$S(u,v) = G(u,v) X(u,v) \quad (3.1)$$

The inverse DFT is then used to obtain the image estimate from its spectrum. The Wiener filter is defined in terms of these spectra:

$H(u,v)$ Fourier transform of the point-spread function (PSF)

$P_s(u,v)$ Power spectrum of the signal process, obtained by taking the Fourier transform of the signal autocorrelation

$P_n(u,v)$ Power spectrum of the noise process, obtained by taking the Fourier transform of the signal autocorrelation.

The Wiener filter is:

$$G(u,v) = \frac{H(u,v)P_s(u,v)}{|H(u,v)|^2 P_s(u,v) + P_n(u,v)} \quad (3.2)$$

Dividing through by P_s makes its behavior easier to explain:

$$G(\mathbf{u}, \mathbf{v}) = \frac{H(\mathbf{u}, \mathbf{v})}{|H(\mathbf{u}, \mathbf{v})|^2 + \frac{P_n(\mathbf{u}, \mathbf{v})}{P_s(\mathbf{u}, \mathbf{v})}} \quad (3.3)$$

The term $\frac{P_n}{P_s}$ can be interpreted as the reciprocal of the signal-to-noise ratio. Where the signal is very strong relative to the noise, $\frac{P_n}{P_s} = 0$ and the Wiener filter becomes $H(\mathbf{u}, \mathbf{v})$ - the inverse filter for the PSF. Where the signal is very weak, $\frac{P_n}{P_s} = \infty$ and $G(\mathbf{u}, \mathbf{v}) = 0$.

For the case of additive white noise and no blurring, the Wiener filter simplifies to:

$$G(\mathbf{u}, \mathbf{v}) = \frac{P_s(\mathbf{u}, \mathbf{v})}{P_s(\mathbf{u}, \mathbf{v}) + \sigma_n^2} \quad (3.4)$$

Where σ_n^2 is the noise variance.

Wiener filters are unable to reconstruct frequency components, which have been degraded by noise. They can only suppress them. Also, Wiener filters are unable to restore components for which $H(\mathbf{u}, \mathbf{v}) = 0$. This means they are unable to undo blurring caused by band limiting of $H(\mathbf{u}, \mathbf{v})$. Such band limiting occurs in any real-world imaging system [46].

3.4.3 Image segmentation

Image segmentation is a branch of digital image processing which focuses on partitioning an image into different parts according to their features and properties. The primary goal of image segmentation is to simplify the image for easier analysis. In image segmentation, you divide an image into various parts that have similar attributes. The parts in which you divide the image are called Image Objects. The kinds of techniques for image segmentation, we can start discussing the specifics. Following are the primary types of image segmentation techniques are thresholding segmentation, edge-based segmentation, region-based segmentation, watershed segmentation, clustering based segmentation Algorithms, and neural networks for segmentation [47].

3.4.3.1 Clustering-Based Segmentation Algorithms

The clustering algorithms is classification algorithms. They are unsupervised algorithms and help you in finding hidden data in the image that might not be visible to a normal vision. This hidden data includes information such as clusters, structures, shadings.

As the name suggests, a clustering algorithm divides the image into clusters (disjoint groups) of pixels that have similar features. It would separate the data elements into clusters where the elements in a cluster are more similar in comparison to the elements present in other clusters [48].

Some of the popular clustering algorithms include fuzzy c-means (FCM), k-means, and improved k-means algorithms. In image segmentation, you would mostly use the k-means clustering algorithm, as it is quite simple and efficient. On the other hand, the FCM algorithm puts the pixels in different classes according to their varying degrees of membership. The most important clustering algorithms for segmentation in image processing K-means clustering.

3.4.3.1.1 K-means Clustering

K-means is a simple unsupervised machine-learning algorithm. It classifies an image through a specific number of clusters. It starts the process by dividing the image space into k pixels that represent k group centroids. Then they assign each object to the group based on the distance between them and the centroid. When the algorithm has assigned all pixels to all the clusters, it can move and reassign the centroids [48].

3.4.4 Feature extraction

Feature extraction is a part of the dimensionality reduction process, in which, an initial set of the raw data is divided and reduced to groups that are more manageable. So when you want to process it will be easier. The most important characteristic of these large data sets is that they have a large number of variables. These variables require a lot of computing resources to process them. These features are easy to process, but still able to describe the actual data set with the accuracy and originality. Image processing is one

of the best and most interesting domain. In this domain basically you will start playing with your images in order to understand them. Statistical methods, structural methods, model-based method and transform based method are used. Texture analysis are rich in visual information and is key component in image analysis [49].

The selection process is to select an optimum subset of features from the enormous set of potentially useful features. Selecting suitable variables is a critical step for successfully implementing an image classification. In medical image analysis, the selection of precise features from among various image modalities is the toughest and most challenging task. Often in data science, there are hundreds or even millions of features and there is a need to create a model that only includes the most important features. This has three benefits. First, make the model simpler to interpret. Second, reduce the variance of the model, and therefore overfitting. Finally, reduce the computational cost (and time) of training a model. The process of identifying only the most relevant features is called "feature selection" [50-51].

3.4.4.1 Texture feature

Texture is a feature used to partition images into regions of interest and to classify those regions, the provides information in the spatial arrangement of colors or intensities in an image the spatial distribution of intensity levels in a neighborhood characterized texture of image. It is a way of describing the spatial distribution of intensities, which makes it useful in classification of similar regions in different images. Statistical methods are extensively used [52].

Haralick texture features are calculated from a Gray Level Co-occurrence Matrix (GLCM), a matrix that counts the co-occurrence of neighboring gray levels in the image. This technique has been widely used in image analysis applications, especially in the biomedical field. It consists of two steps for feature extraction. The GLCM is a square matrix that has the dimension of the number of gray levels N in the region of interest (ROI). Each texture feature is a function of the elements of the GLCM, and represents a specific relation between neighboring voxels. The GLCM is computed in the first step, while the texture features based on the GLCM are calculated in the second step [53].

A co-occurrence matrix is a two-dimensional array, P , in which both the rows and the columns represent a set of possible image values. A GLCM $p_d[i,j]$ is defined by first specifying a displacement vector $d=(dx,dy)$ and counting all pairs of pixels separated by d having gray levels I and j . The GLCM is defined by:

$$p_d[i,j] = n_{ij} \quad (3.5)$$

Where n_{ij} is the number of occurrences of the pixel values (i,j) lying at distance d in the image. The co-occurrence matrix p_d has dimension $n \times n$, where n is the number of gray levels in the image. The texture features can indicate:

3.4.4.1.1 Contrast

Is a measure of intensity or gray level variations between the reference pixel and its neighbor. The contrast size reflects the sharpness, texture density and the depth extent of groove. The visual effect is clearer if image has deeper groove and bigger contrast. If the element values, which are far away from the diagonal, are bigger in GLCM, the contrast value will be bigger large contrast reflects large intensity differences in GLCM.

3.4.4.1.2 Homogeneity

Homogeneity measure the similarity of pixels. A diagonal gray level co-occurrence matrix gives homogeneity of one. It becomes large if local textures only have minimal changes. A homogeneous image will result in a co-occurrence matrix with a combination of high and low $P[i,j]$'s. Where the range of gray levels is small the $P[i,j]$ will tend to be clustered around the main diagonal. A heterogeneous image will result in an even spread of $P[i,j]$'s.

3.4.4.1.3 Entropy

Entropy is a measure of information content. It measures the randomness of intensity distribution. The equation of entropy such a matrix corresponds to an image in which there are no preferred gray level pairs for the distance vector d . Entropy is highest when all entries in $P[i,j]$ are of similar magnitude, and small when the entries in $P[i,j]$ are unequal.

3.4.4.1.4 Correlation

Correlation is a measure of image linearity. It is the measure of gray tone linear dependencies in the image. Feature values range from minus one to 1, these extremes indicating perfect negative and positive correlation respectively. Correlation will be high if an image contains a considerable amount of linear structure [54].

3.4.4.1.5 Autocorrelation

It is a mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals [55].

3.4.4.1.6 Energy

It is derived from the Angular Second Moment (ASM). The ASM measures the local uniformity of the gray levels. When energy equals to one, the image is believed to be a constant image.

3.4.4.1.7 Variance

Increasing weight given to greater gray value differences.

3.4.4.1.8 Sum average

Average sum of gray levels.

3.4.4.1.9 Sum of variance

Variance of sum of gray levels

3.4.4.1.10 Sum of entropy

Uniform (flat) distribution of sum of gray levels has maximum entropy.

3.4.4.1.11 Difference entropy

Uniform (flat) distribution of difference of gray levels has maximum entropy.

3.4.4.1.12 Difference variance

Variance of difference of gray levels

3.4.4.1.13 Information measure of correlation 1

Normalized mutual information.

3.4.4.1.14 Information measure of correlation 2

Difference between joint entropy and joint entropy assuming independence [56].

Other feature the tow shape descriptors were calculated like (area and solidity) corresponding to the object's physical dimensional measures (size, position, and shape) that characterize its appearance.

3.4.4.1 Equations of Texture feature

This table (3.1) below the left column shows the original definitions, and the right column shows the modifications needed to make the features invariant to the number of gray-levels. There was an error in the definition of Sum variance in Haralick *et al.*, which has been corrected. $\lambda_2(Q(i, j))$ denotes the second largest eigenvalue of a matrix $Q(i, j)$. Note, however, that this feature is computationally unstable, and was therefore not included in the examples in this work. For symmetric GLCMs, μ_x and μ_y is identical, and is represented by μ in the expression for Cluster prominence and Cluster shade [57].

Table (3.1): Equations of Texture feature

Feature	Original expression
Autocorrelation	$\sum_{i=1}^N \sum_{j=1}^N (i \cdot j) p(i, j)$
Cluster prominence	$\sum_{i=1}^N \sum_{j=1}^N (i + j - 2\mu)^3 p(i, j)$
Cluster shade	$\sum_{i=1}^N \sum_{j=1}^N (i + j - 2\mu)^4 p(i, j)$
Contrast	$\sum_{i=1}^N \sum_{j=1}^N (i - j)^2 p(i, j)$
Correlation	$\sum_{i=1}^N \sum_{j=1}^N \left(\frac{i - \mu_x}{\sigma_x} \right) \left(\frac{j - \mu_y}{\sigma_y} \right) p(i, j)$
Difference entropy	$-\sum_{k=0}^{N-1} p_{x-y}(k) \log p_{x-y}(k)$
Difference variance	$\sum_{k=0}^{N-1} (k - \mu_{x-y})^2 p_{x-y}(k)$
Dissimilarity	$\sum_{i=1}^N \sum_{j=1}^N i - j \cdot p(i, j)$
Energy	$\sum_{i=1}^N \sum_{j=1}^N p(i, j)^2$
Entropy	$-\sum_{i=1}^N \sum_{j=1}^N p(i, j) \log p(i, j)$
Homogeneity	$\sum_{i=1}^N \sum_{j=1}^N \frac{p(i, j)}{1 + (i - j)^2}$
Information measure of correlation 1	$\frac{HXY - HXY1}{\max(HX, HY)}$
Information measure of correlation 2	$\sqrt{1 - \exp[-2(HXY2 - HXY)]}$
Inverse difference	$\sum_{i=1}^N \sum_{j=1}^N \frac{p(i, j)}{1 + i - j }$
Maximum probability	$\max_{i,j} p(i, j)$
Sum average	$\sum_{k=2}^{2N} k p_{x+y}(k)$
Sum entropy	$-\sum_{k=2}^{2N} p_{x+y}(k) \log p_{x+y}(k)$
Sum of squares	$\sum_{i=1}^N \sum_{j=1}^N (i - \mu)^2 p(i, j)$
Sum variance	$\sum_{k=2}^{2N} (k - \mu_{x+y})^2 p_{x+y}(k)$
Maximal Correlation Coefficient	$\sqrt{\lambda_2(Q(i, j))}$

3.4.4 Classification

Classification tasks are simply related with predicting a category of a data (discrete variables). Some of the common use cases could be found in the area of healthcare such as whether a person is suffering from a particular disease or not. The ML methods such as following could be applied to solve classification tasks [58]:

- Kernel discriminant analysis (Higher accuracy)

- K-Nearest Neighbors (Higher accuracy)
- Artificial neural networks (ANN) (Higher accuracy)
- Support vector machine (SVM) (Higher accuracy)
- **Random forests (Higher accuracy)**
- Decision trees
- Boosted trees
- Logistic regression
- Naive Bayes
- Deep learning

3.4.4.1 Random Forests

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [59].

A Random Forest Classifier consists of a collection of decision tree classifiers where the decision trees are independently distributed random vectors and each tree casts a unit vote for the most popular class for a given input. Random Forests' main impact was on the analysis tasks that required understanding spatial context within the images. We take a specific angle and view Random Forests as a machine-learning tool that can integrate contextual information. We position the algorithm and its contributions within the larger field from this respect. Lastly, we briefly discuss how Random Forests and deep learning methods relate to each other and how they differ. The average of many trees is not sensitive to noise as opposed to a single tree which is highly sensitive to noise [60].

3.4.4.2 Random forest algorithm

The random forest algorithm can split into two stages [61]:

- Random forest creation.
- Perform prediction from the created random forest classifier.

3.4.4.2.1 Random forest creation

The training process of random forests can be described by algorithm as follow:

1. Randomly select k features from total m features.

Where $k \ll m$

2. Among the k features, calculate the node d using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1 to 3 steps until l number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for n number times to create n number of trees.

The beginning of random forest algorithm starts with randomly selecting k features out of total m features. Then, using the randomly selected k features to find the root node by using the best split approach. The next stage, calculating the daughter nodes using the same best split approach. Then repeating the first 3 stages until form the tree with a root node and having the target as the leaf node. Finally, repeat 1 to 4 stages to create n randomly created trees. This randomly created trees form the random forest [61].

3.4.4.2.1 Random forest prediction

The following algorithm describes the testing process of the random forests:

1. Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target).
2. Calculate the votes for each predicted target.
3. Consider the high voted predicted target as the final prediction from the random forest algorithm.

To perform the prediction using the trained random forest algorithm. Need to pass the test features through the rules of each randomly created tree. Each random forest will predict different target (outcome) for the same test feature. Then by considering each predicted target, votes will be calculated. This concept of voting is known as majority voting [61].

3.4.5 Background on Matlab and the Image Processing Toolbox

The name MATLAB stands for matrix laboratory. MATLAB was written originally to provide easy access to matrix software developed by the LINPACK (Linear System Package) and EISPACK (Eigen System Package) projects. Today, MATLAB engines incorporate the LAPACK (Linear Algebra Package) and BLAS (Basic Linear Algebra Subprograms) libraries, constituting the state of the art in software for matrix computation. It has powerful built-in routines that enable a very wide variety of computations. It also has easy-to-use graphics commands that make the visualization of results immediately available. Specific applications are collected in packages referred to as toolbox. There are toolboxes for signal processing, symbolic computation, control theory, simulation, optimization, and several other fields of applied science and engineering [62]

MATLAB is a high performance language for technical computing .it is used in graphical user interface building. Computer-aided diagnosis (CAD) is abroad concept that integrates image processing, computer vision, mathematics, physics, and statistics into computerized techniques that assist radiologists in their medical decision- making processes [63].

CHAPTER FOUR

METHODOLOGY

4.1 Introduction

This chapter discusses the method approached to building classification system for COVID-19 based on random forests, which basic consist from five stages as shown in figure (4.1) below:



Figure (4.1): block diagram of research methodology.

4.1.1 Data collection

The image dataset used here is the Computed Tomography (CT) images of the chest. The dataset is obtained from the GitHub. The datasets consists 794 images divided into 395 positive COVID-19 cases and 379 negative COVID-19 cases. These images are donated by the various hospitals and collected from different sources and are of high quality but the hospital refuse to share the patient's data.

In this collection, the specification of CT data vary in name the rename of all images is COVID and non-COVID, then sort images by number with name in file. If images is not sort the result is error, also vary in sizes, quantized (8 bits) and format (jpg, png, jpeg). The images in this research RGB (red, green, and blue) image correctly, whether its class is 8-bit unsigned integer arrays (uint8).

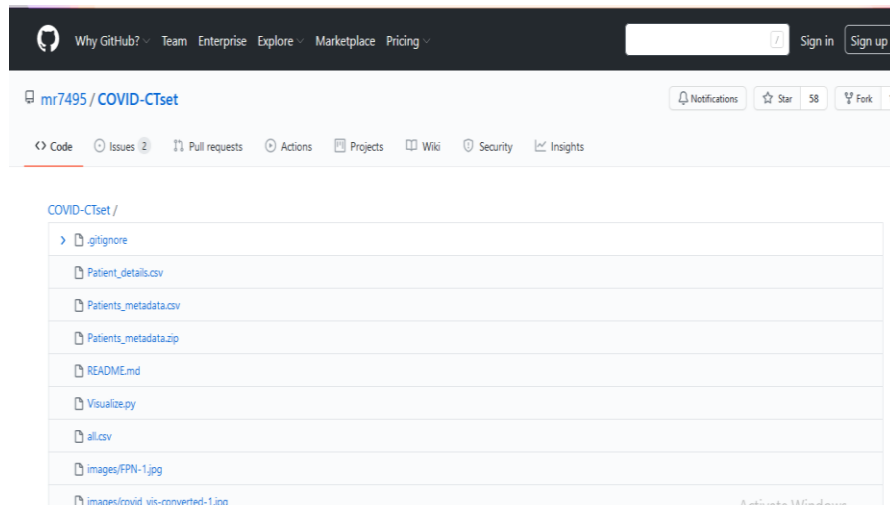


Fig. 4.1 the GitHub website.

4.1.2 Image processing

Preprocessing is first stage in the begging process to improve the quality of image and better resolution. Preprocessing steps including contrast enhancement and applied wiener filter.

First step is converted input image to gray scale to making them easier to segment and identify. The type of noise in our data is speckle Noise, which we found it in literature studies. There are many de- noising filters could be used. We calculate the MSE, SNR and PSNR for different images after applying different filters (Gaussian, Adaptive median, Wiener, and Median) to compare between them and select the best choice, which gives us the best result and that which we found it in literature studies.

Winer filter was found to be most compatible method to remove speckle noise and blurred from the images. It gives the better results than the other filters of preserving useful details in an image in this research. The filtered image is used, as the input for next step is image segmentation.

4.1.3 Image segmentation

This process is the most important in the analysis of image. The main purpose of this step is to find region of interest (ROI) by learning the homogeneity feature of the region, in other words to divide an image into regions that have a strong correlation with objects or areas of the real world contained in the image.

The segmentation process of lung image is extract alveoli after input image use morphological reconstruction to constructing an image from small components and smooth image of all particle and speckle noise. The reconstruction process occurs on the marker image, which is created by applying dilations or erosions on the lung image. The second step of extract image use threshold of fifty percentage of all number of pixels to compare the mean pixel with pixel of lung image to separate the high intensity from low intensity of image and remove any speckle noise. After that apply global threshold use Otsu's method, which chooses the threshold to minimize the interclass variance of the black and white pixels and convert image from black to white of any pixel and invert this process. The clear the image border to reduce the overall intensity level in addition to suppressing border structures and show of largest tow objects of lung image (this left and right lung) then fill image regions and holes.

This process of lung mask apply clustering texture (K-means Clustering) but the objective of segment the alveoli of lung using clustering image using three number of group intensity (the low intensity(black), medium intensity(gray) and high intensity(white)). To show only alveoli of lung determine the group three (low intensity) inside lung mask, then after this process some noise to remove this noise use the morphological filter image, retaining only those objects with areas between ten to infinity by experimentally. Sometimes not found alveoli or branches not apply extraction of image the result of image mask is empty, in other word just image is black (normal) otherwise this image is probable subjected object (abnormal) for next step feature extraction.

4.1.4 Feature extraction

Feature extraction is process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. A critical requirement for alveoli of lung is to extract quantitative features information. A total of 24 features was used in one alveoli comprising two major categories were extracted; 24 statistical texture features, 2 shape features.

A second order statistical technique (Harlick texture feature) was proposed to extract features from segmented and normalized image described by statistics of distribution of gray level or intensities in the texture.

Gray Level Co-occurrence Matrix (GLCM) was used to compute the twenty tow Harlick features in 0 degree angle; energy, correlation, entropy, homogeneity, maximum probability, variance sum average ,sum variance ,sum entropy ,difference variance ,difference entropy, information measure of correlation1 , and information measure of correlation2. Some other texture features like autocorrelation, contrast, correlation, cluster prominence, cluster shade, dissimilarity, inverse difference (INV) is homogeneity, Inverse difference normalized (INN) ,inverse difference moment normalized were obtained.

The tow shape descriptors were calculated like (area and solidity) corresponding to the object's physical dimensional measures (size, position, and shape) that characterize its appearance. All this feature the result tow values then mean were calculated.

4.1.5 Features Selection

Random Forests are often used for feature selection in a data science workflow. The reason is because the tree-based strategies used by random forests naturally ranks by how well they improve the purity of the node. This mean decrease in impurity over all trees (called gini impurity). Nodes with the greatest decrease in impurity happen at the start of the trees, while nodes with the least decrease in impurity occur at the end of trees. Thus, by pruning trees below a particular node, we can create a subset of the most important features.

4.1.6 Classification and Evaluation

The Classification for lung CT images aims at classifying the lung into (normal and abnormal) using the selected extracted features of Haralick this is done by using

Random Forests Classification technique. The Random Forest predicts the output depending on a training set with observations' and their corresponding targets.

4.1.6.1 Random Forest steps

- Create the Data.
- Split the Data into Training and Test Sets.
- Train a Random Forest Classifier.
- Identify and Select Most Important Features.
- Create a Data Subset with Only the Most Important Features.
- Train a New Random Forest Classifier Using Only Most Important Features.
- Compare the Accuracy of the Full Feature Classifier To the Limited Feature Classifier.

The extracted features from all the dataset was gathered in on Microsoft Excel file. The data was split into training and testing set. The training size is 75%, and the testing is 25%. The Random forest model was built using reduce a predictor set. Because prediction time increases with the number of predictors in random forests, a good practice is to create a model using as few predictors as possible. Grow a random forest classifier of 200 regression trees according the best two predictors only using Predictor Selection Character vector specifying the algorithm for choosing the best split predictor of interaction curvature the split predictor is chosen by minimizing the p-value of a chi-square test of independence between each predictor and response and minimizing the p-value of a chi-square test of independence between each pair of predictors and response. The other determine Surrogate is decision tree finds 10 surrogate splits at each branch node.

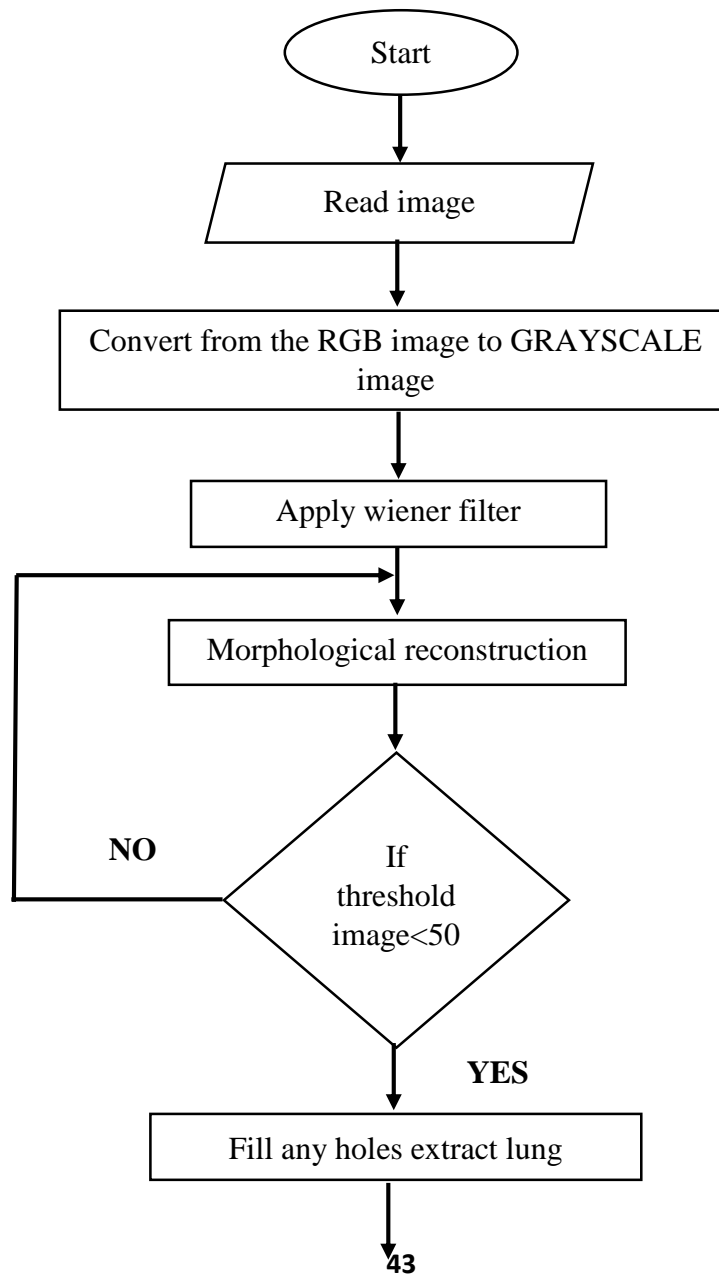
When set to an integer, decision tree finds at most the specified number of surrogate splits at each branch node. Use surrogate splits to improve the tree accuracy for data with missing values or to compute measures of association between predictors. Then determine maximal number of decision splits per tree is 500. Finally was testing this tree by predict and calculate the rate of test all this process is train of random forest model.

In the implementation, the classifier consists from decision tree and the data consists from 794 images and 24 features. Due to the limited data used in the classification process so that the data cannot be divided into a training group and a test

group, the k folding method was used to solve this problem and to evaluate the performance of the system. For determining the image covid or non-covid of all features in one image, then to compare between feature and class already given if equal the classify image is true. For evaluating the model, the predicted result of the RF model using the testing set was, then constructed on confusion matrix in order to find performance; by calculating the accuracy, sensitivity, specificity, and error rate.

4.1.7 Flow chart algorithm

The figure (4.2) below explain the code of MATLAB the all process and steps the research



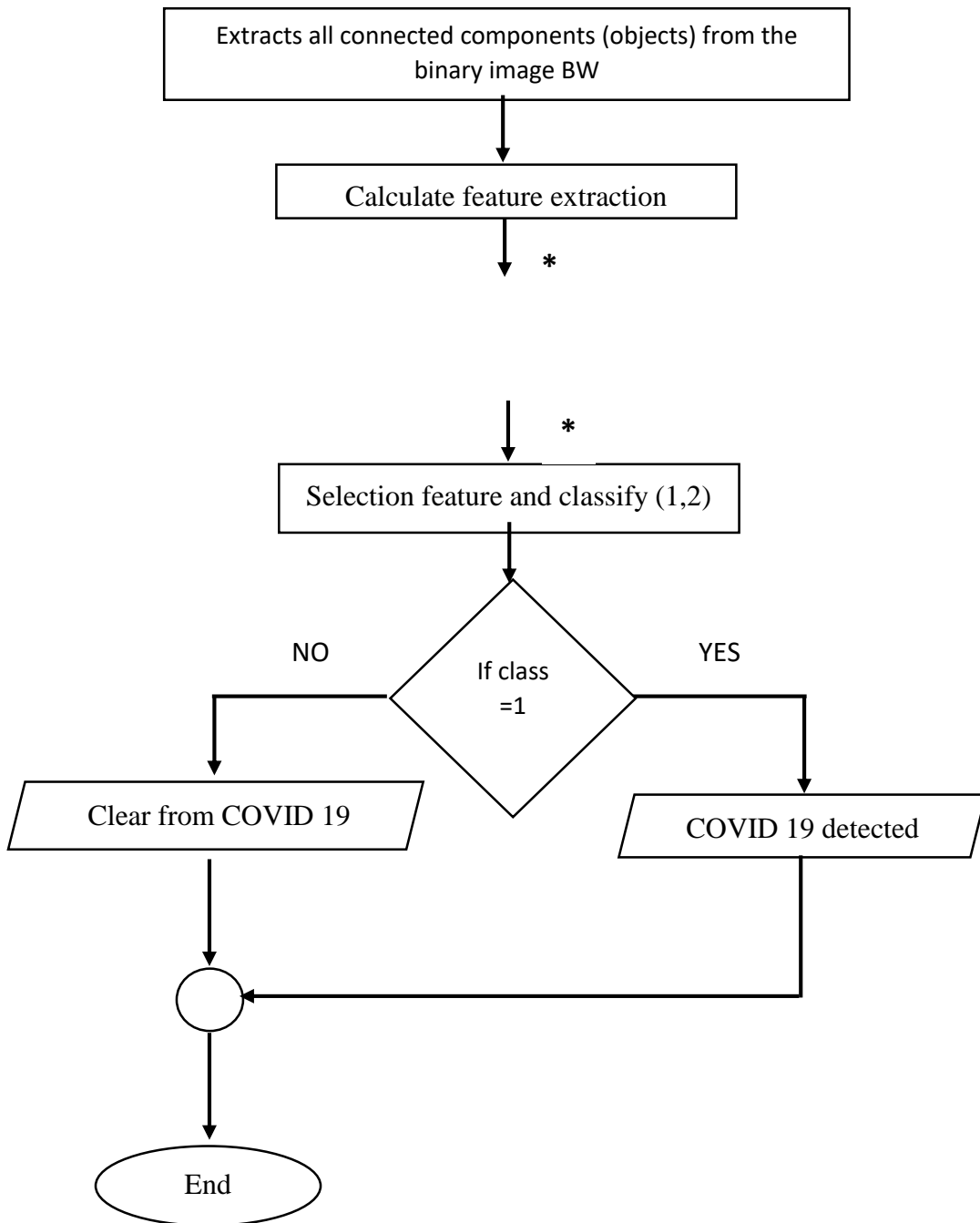


Figure (4.2): Flow chart of algorithm

CHAPTER FIVE

RESULT AND DISCUSSION

5.1 Result

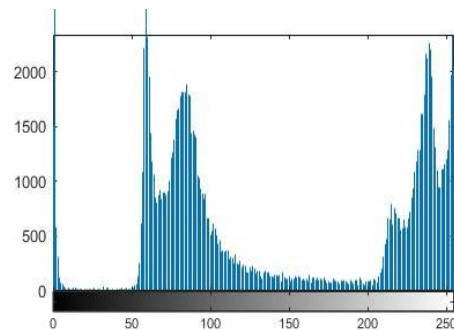
In this chapter, based on the collected data and applying methods and condition described in the methodology chapter the following results were obtained. A collection of 794 of CT images were using for training and testing the system.

5.1.1 Result of image preprocessing of non-COVID image

In preprocessing stage, the figure (5.1) below show the original image (gray scale) and histogram after and before applying the pre-processing step of non-COVID image to show and validate of the information of image.



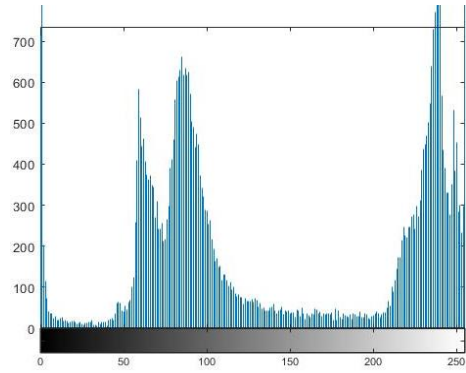
(A) Gray scale



(b) Histogram for the gray scale image



(C) Applying wiener filter

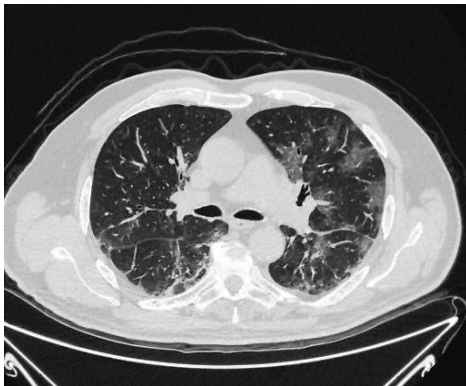


(D) Histogram of (C) image

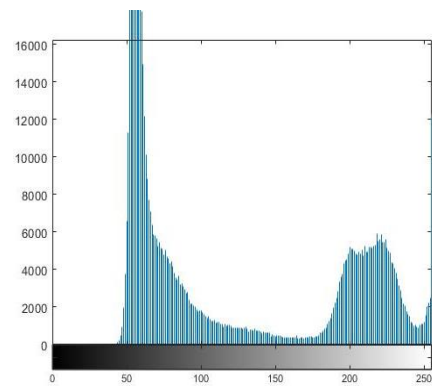
Figures (5.1) showing Image preprocessing of non-COVID

5.1.2 Result of preprocessing of COVID 19 image

In preprocessing stage, the figure (5.2) below show the original image (gray scale) and histogram after and before applying the pre-processing step of COVID 19 to show and validate of the information of image..



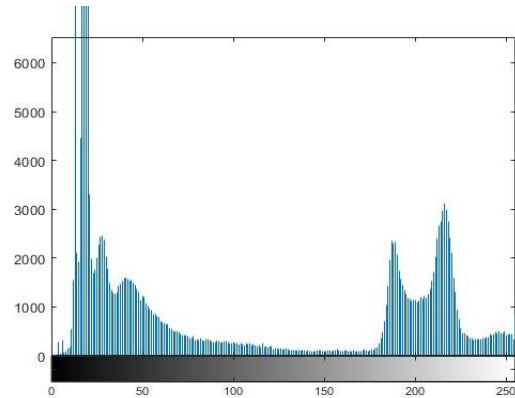
(A) Gray scale image



(b) Histogram for the gray scale image



(C) Applying wiener filter



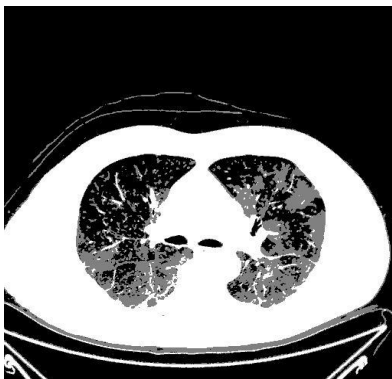
(D) Histogram of (C) image

Figures (5.2) showing Image preprocessing of non-COVID 19

In these step, after it have tested and converted the original image into gray scale as in image (A), we were shows the histogram of the image in (B) to follow the change in image in next steps. The second step the wiener filter is used to remove speckle noise from the lung CT original image. It is suitable because the data is downloaded from the website and it is already enhancement, as shown in image after filtering (C) and show the histogram of it (D).the figures(5.1) are COVID 19 and Figures (5.2) are non-COVID 19.

5.1.3 Results of Segmentation

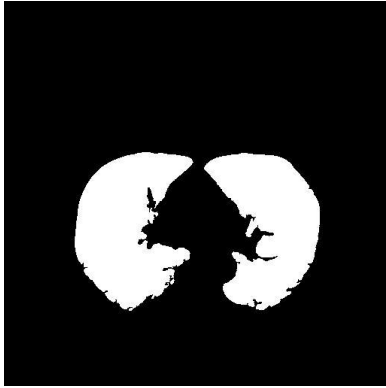
The resulted image after following the segmentation process are shown in the figures (5.4) below



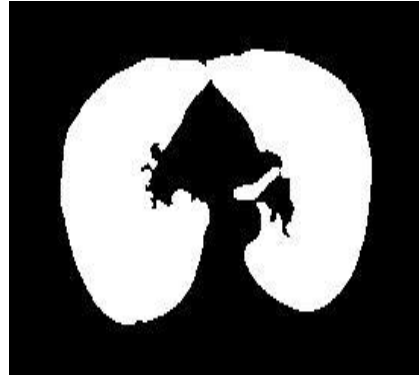
(a)K-mean of COVID 19 image



(a) K-mean of nodels image



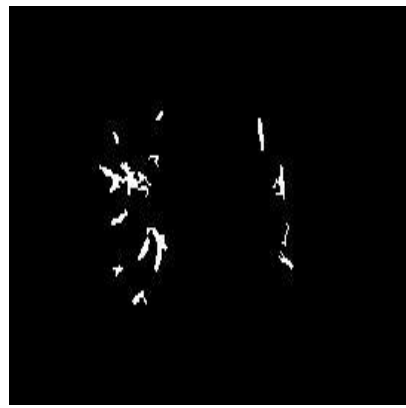
(b) Lung mask of COVID 19 image



(b) lung mask of nodels image



(c) COVID mask



(c) Nodels segmented

Figures (5.3) Shows the Segmentation stage

This stage shows the Segmentation of nodules or COVID 19 in two steps as explained in methodology. First, step (A) as we mentioned before is segment the nodules or COVID 19 by Applying K-means Clustering this detection step was performed to make sure only overlapping nodules are separate to reduce error. Image (B) is after closing operation in lung mask to fill the gaps, this carried out by manual selection with predefined ranges using properties of each component in the binary image to select large object .image C by using area of alveoli remove all object of image and determine COVID 19 mask or nodules and it could use it as input in the classification.

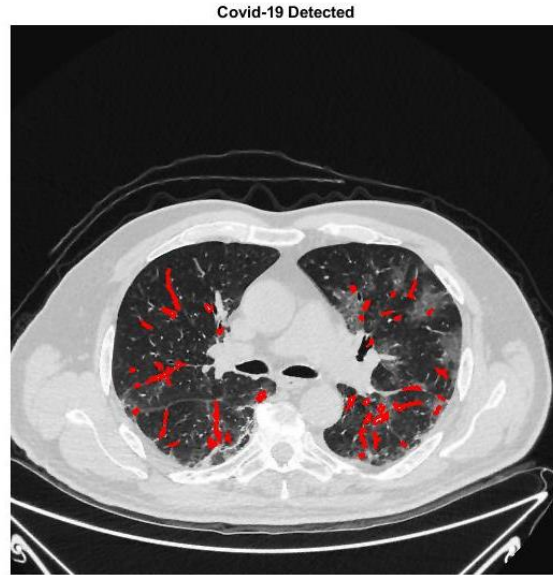


Figure (5.4) Image after diagnosis the red color explain the place of COVID 19.

5.1.4 Results of feature extraction and selection

Features were extracted by applying Haralick (GLCM) and shape features code to all 794 images “combination of 24 features” any image contain about 30 value of on feature. Using features selection were selected 19 features and 100 values for each feature because the number of values larg, then arranged in an excel sheet in a form of matrix and fed it as input for the random forest. Results shown in figures bellow:

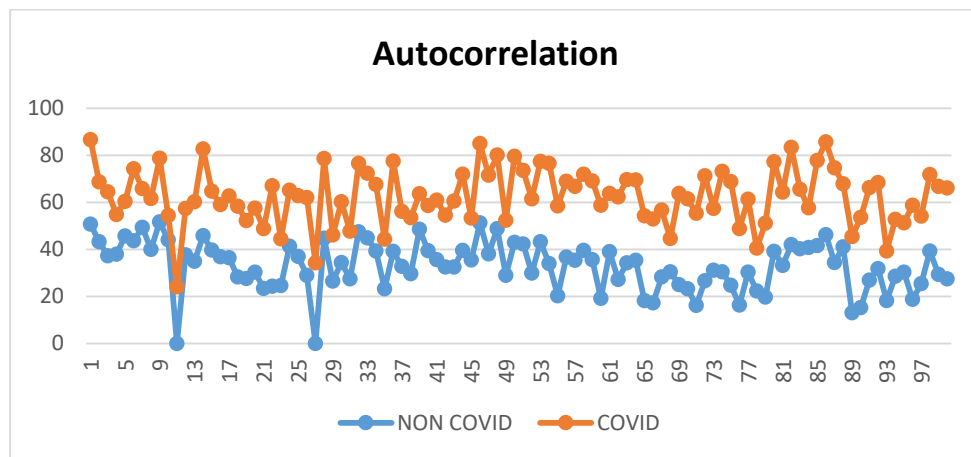


Figure 5.5 Autocorrelation feature.

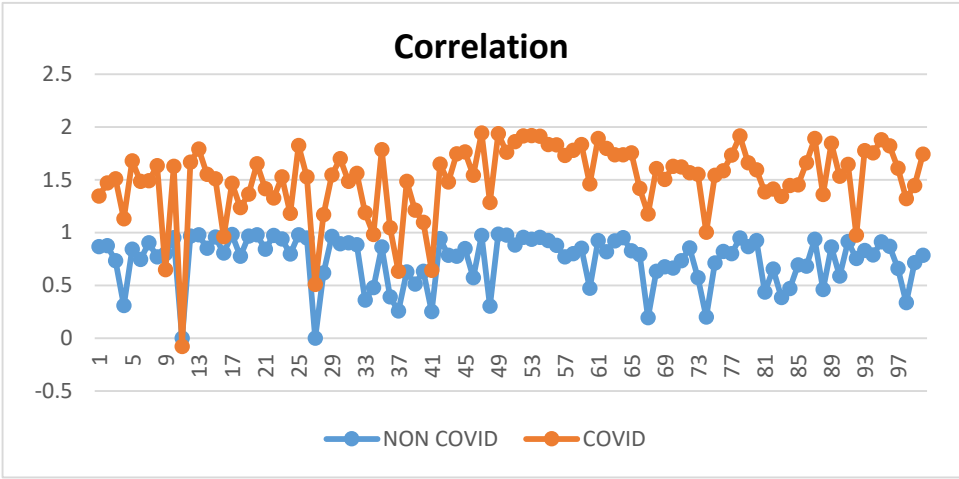
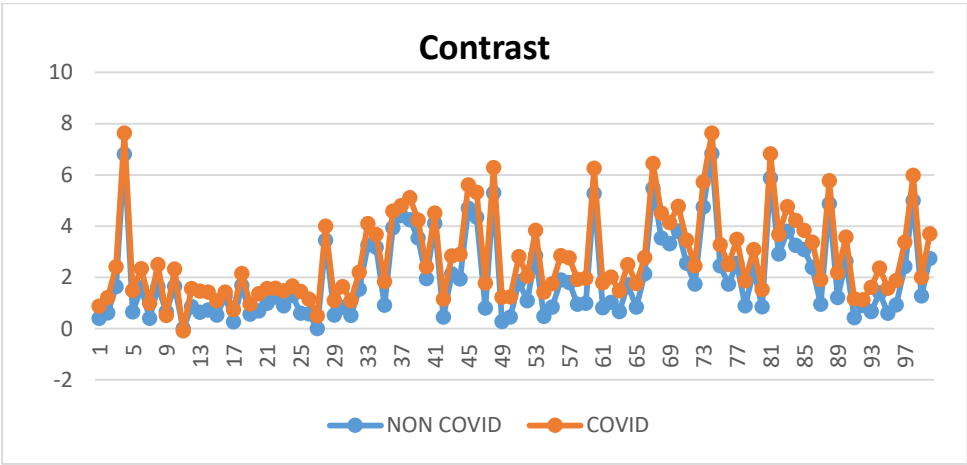


Figure (5.6) contrast feature.

Figure (5.7) Correlation feature.

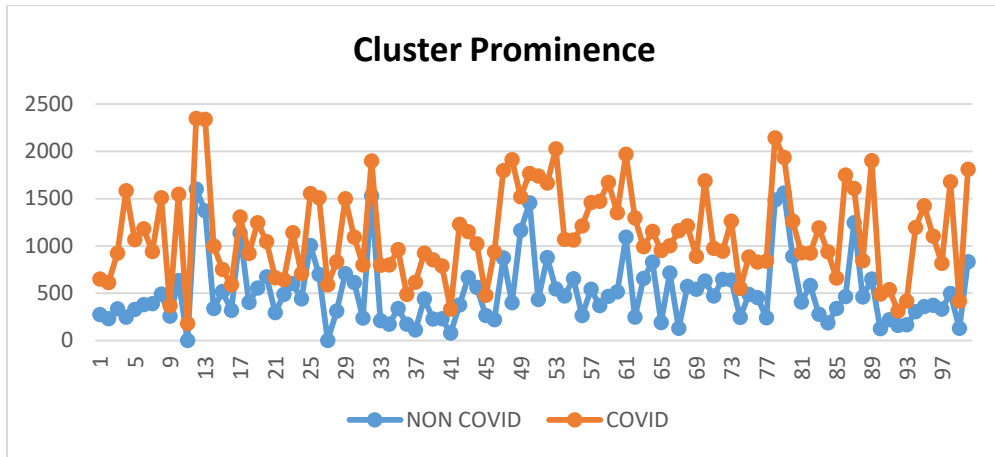


Figure (5.8) Cluster Prominence feature.

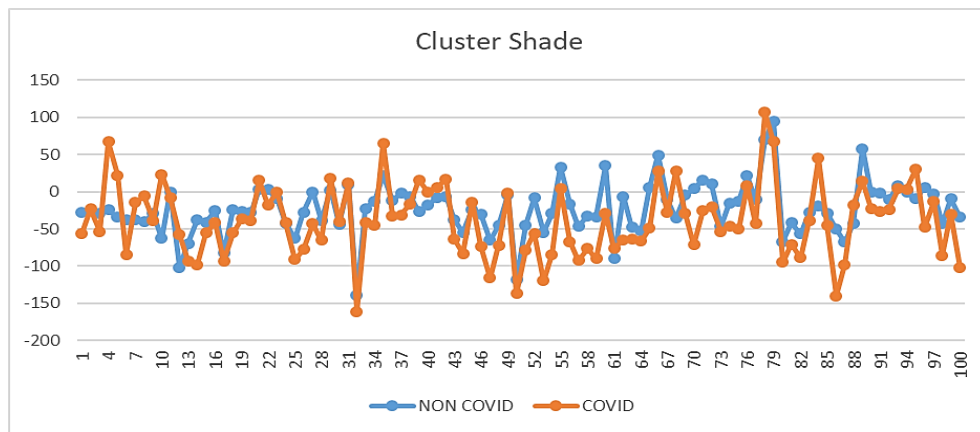


Figure (5.9) cluster shade feature.

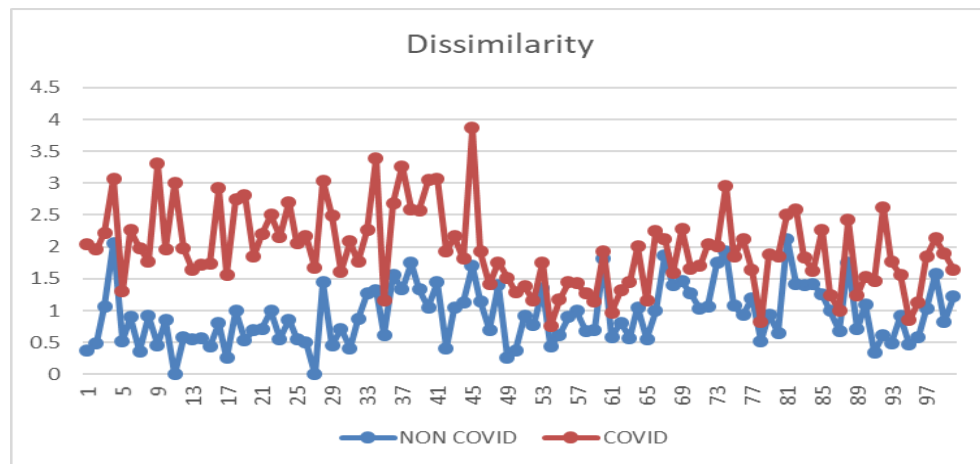


Figure (5.10) dissimilarity feature.

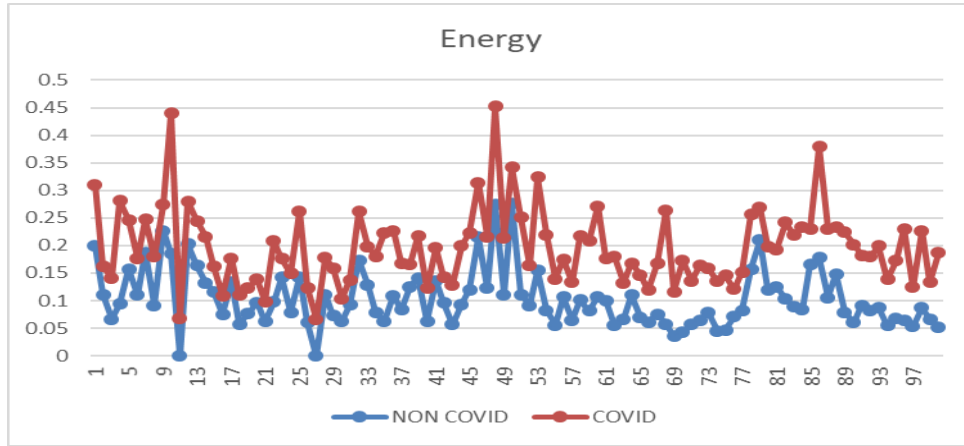


Figure (5.10) Energy feature.

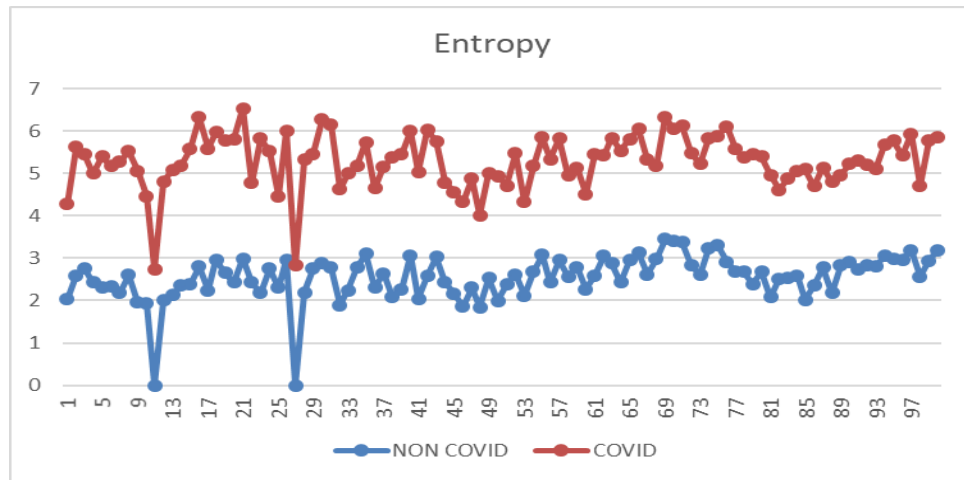


Figure (5.12) Entropy feature.

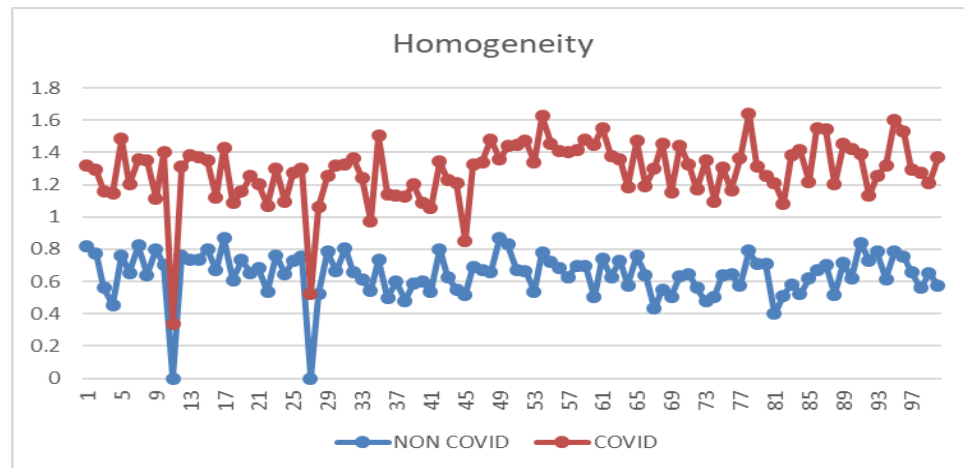


Figure (5.13) Homogeneity feature.

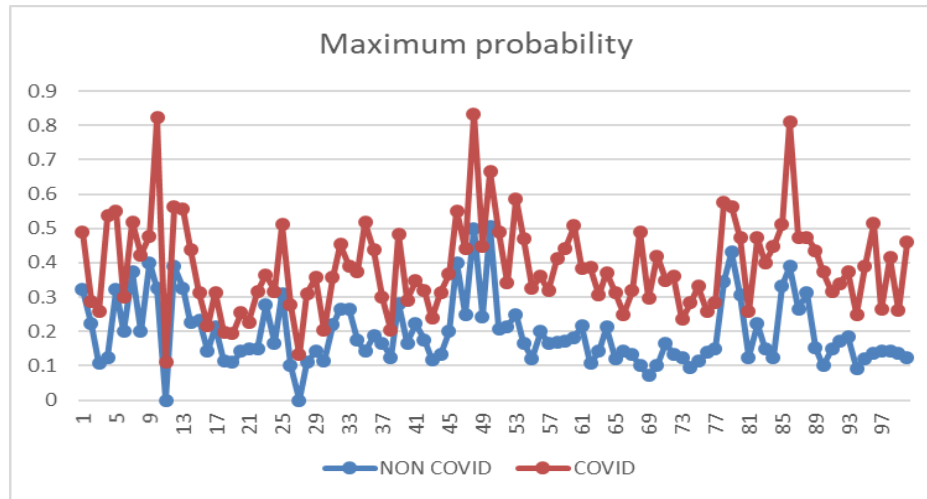


Figure (5.14) Maximum probability feature.

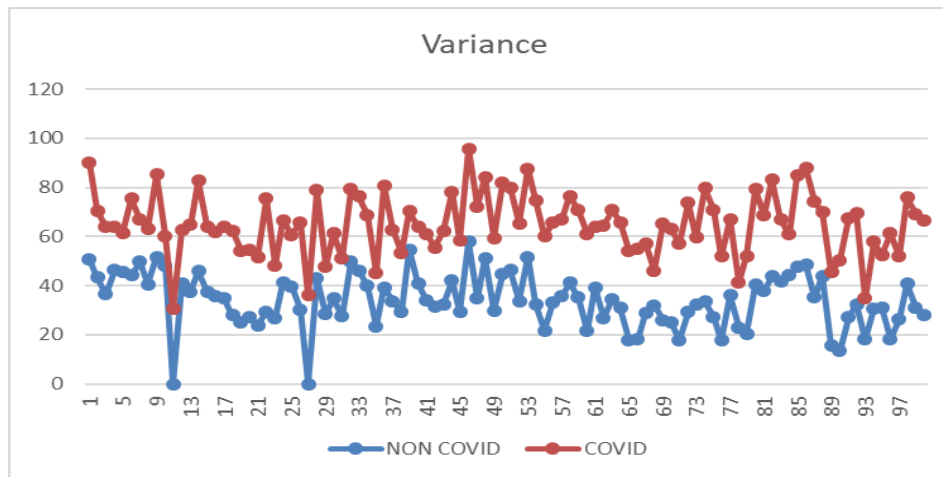


Figure (5.15) Variance feature.

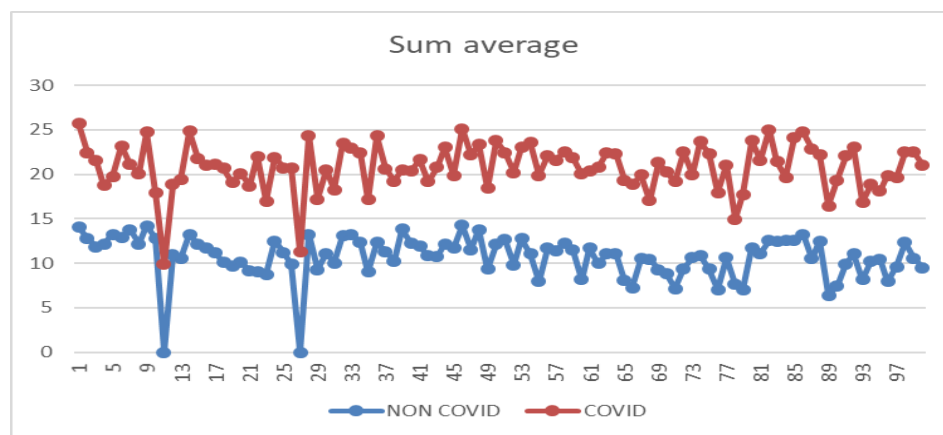


Figure (5.16) sum average

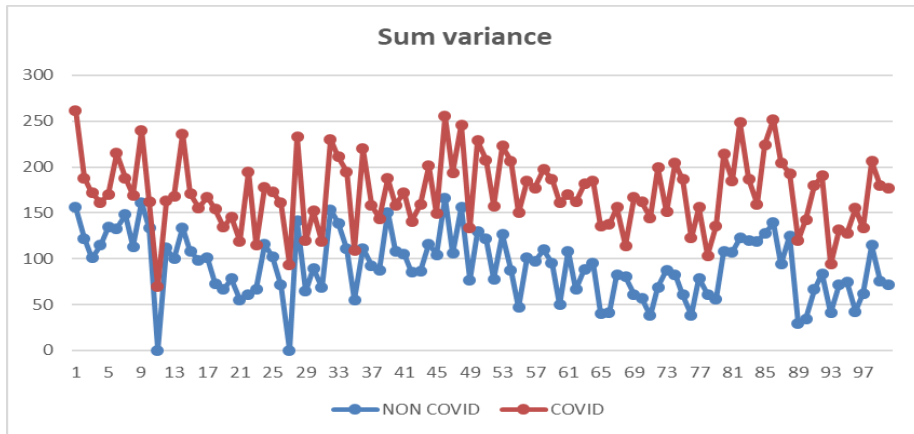


Figure (5.17) Sum variance feature.

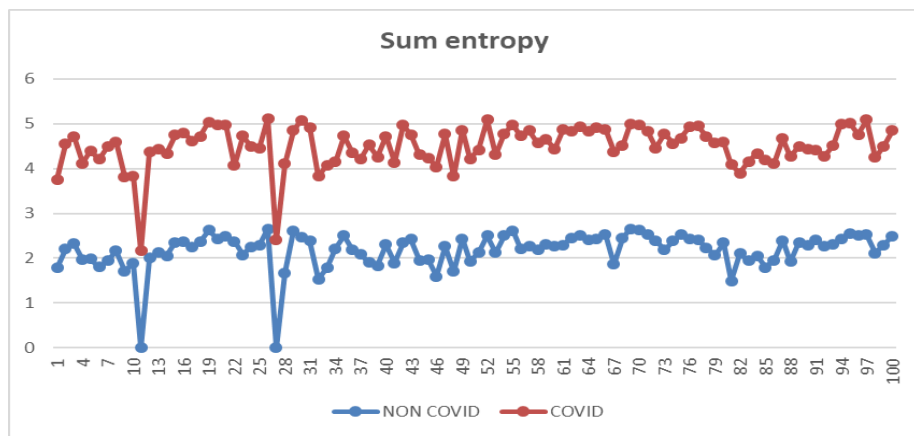


Figure (5.18) Sum entropy feature.

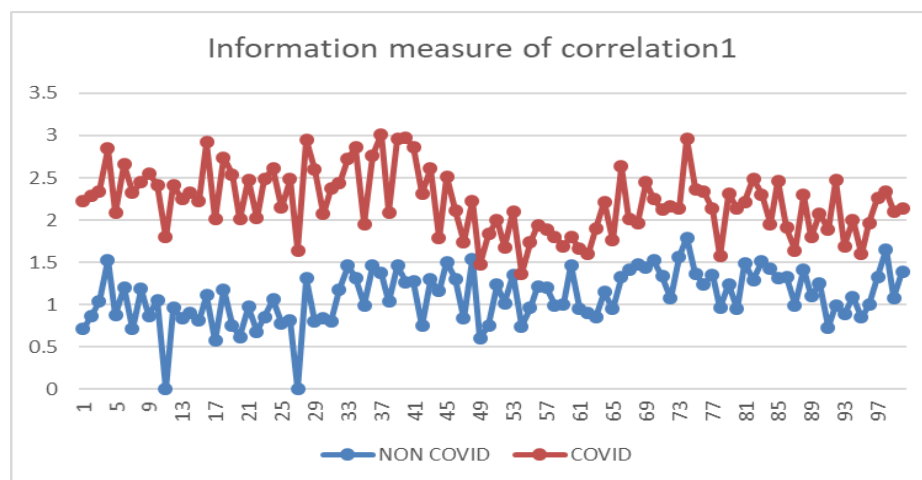


Figure (5.19) Information measure of correlation1 feature.

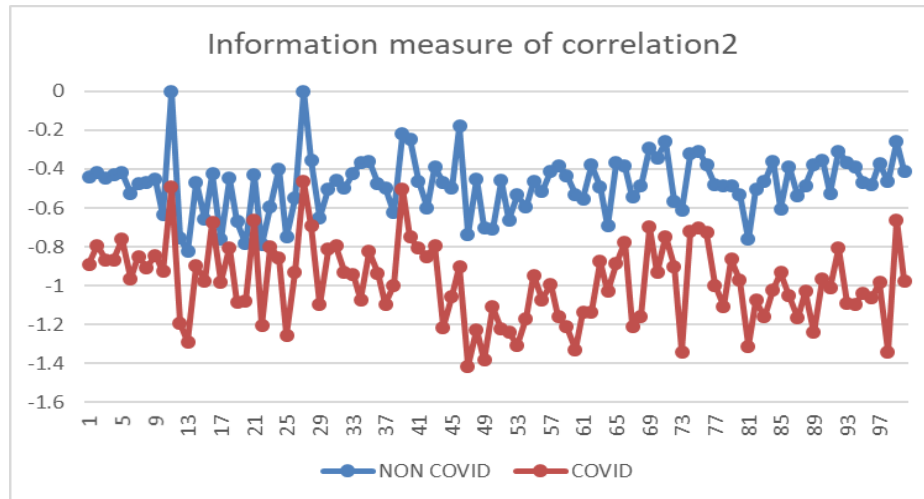


Figure (5.20) Information measure of correlation 2 feature.

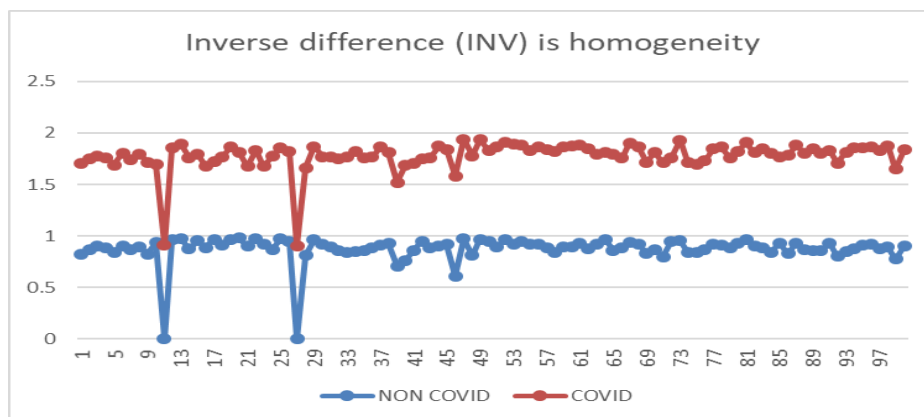


Figure (5.21) Inverse difference (INV) feature.

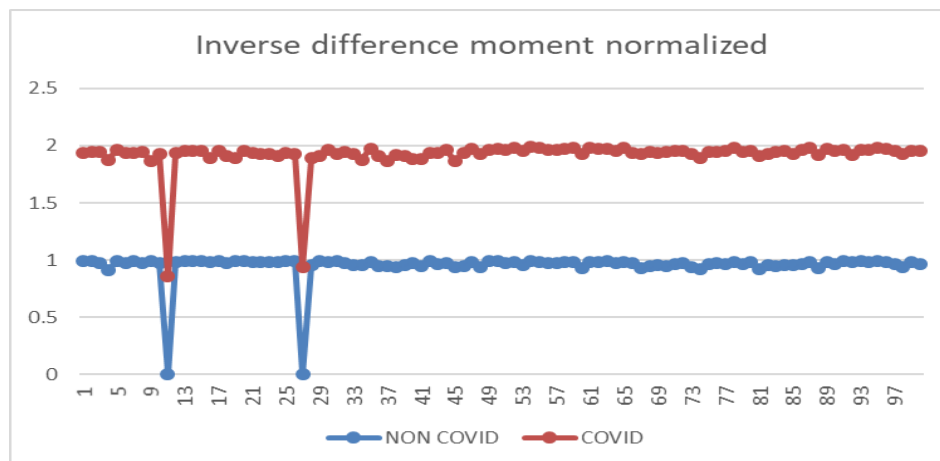


Figure (5.22) Inverse difference normalized feature.

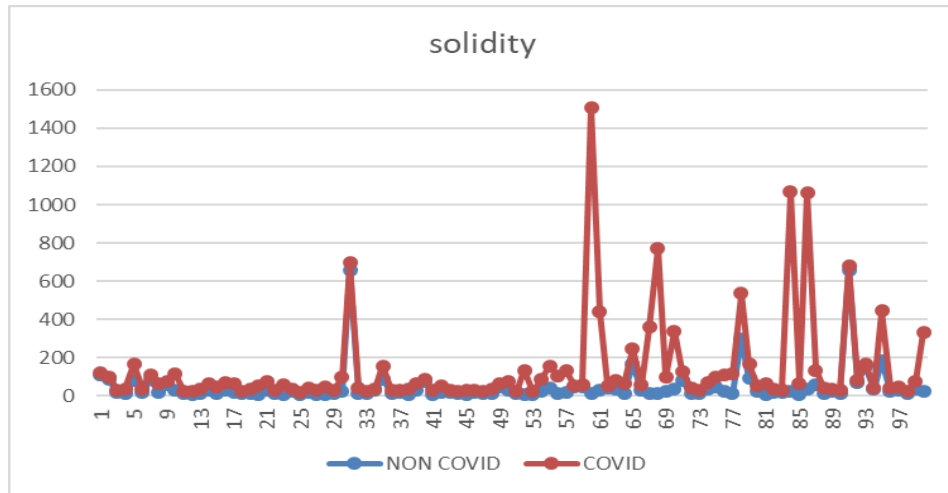


Figure (5.23) solidity feature.

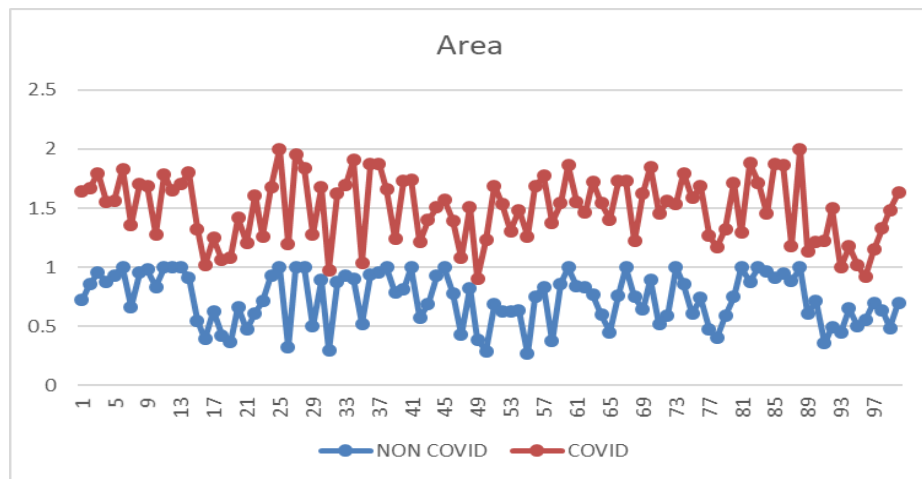
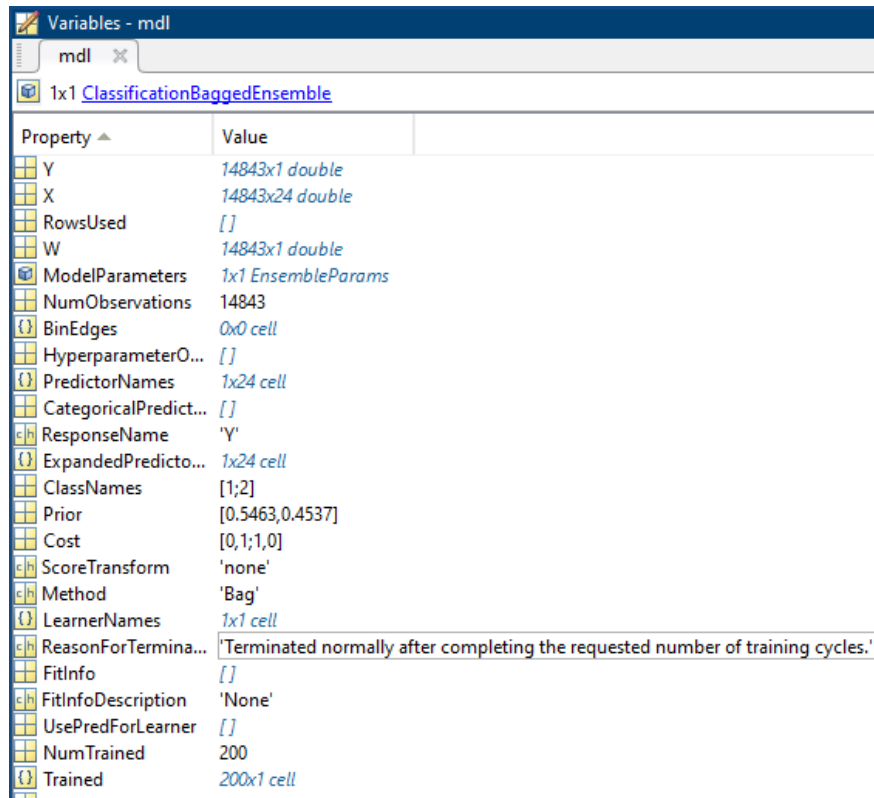


Figure (5.24) Area feature.

These figures as we mentioned shows the result of selected features, which are critically, could separate our data into the two classes (NON COVID and COVID). In each figure above there is a representation of the relationship between the number of sample values of image (X-axis) and specific selected features (Y-axis), Each line shows variable values of samples and the correspond value of feature the other value of feature in appendix.

5.1.5 Result of classification

The random forests method does not need to select the most important features because it arranges the features automatically, the most important features place at the beginning of the decision tree and the less important features at the bottom of the tree .so it used in the process of features engineering.



Property	Value
Y	14843x1 double
X	14843x24 double
RowsUsed	[]
W	14843x1 double
ModelParameters	1x1 EnsembleParams
NumObservations	14843
BinEdges	0x0 cell
HyperparameterO...	[]
PredictorNames	1x24 cell
CategoricalPredict...	[]
ResponseName	'Y'
ExpandedPredicto...	1x24 cell
ClassNames	[1;2]
Prior	[0.5463,0.4537]
Cost	[0,1;1,0]
ScoreTransform	'none'
Method	'Bag'
LearnerNames	1x1 cell
ReasonForTermina...	'Terminated normally after completing the requested number of training cycles.'
FitInfo	[]
FitInfoDescription	'None'
UsePredForLearner	[]
NumTrained	200
Trained	200x1 cell

Figure (5.25): specifications of the random forest classifier.

5.1.5.1 Performance measures

- **Sensitivity (SE)**

The ability of a test to correctly identify those with the disease (true positive rate).

$$SE=TP/ (TP+FN) \quad (5.1)$$

- **Specificity (SP)**

The ability of the test to correctly identify those without the disease (true negative rate).

$$SP=TN/ (TN+FP) \quad (5.2)$$

- **Accuracy**

The presents all samples correctly classified.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (5.3)$$

Correct Rate = 97.254

Sensitivity = 98.837

Specificity = 95.690

Error Rate = 2.746

Figure (5.26) Result of performance training random forest

Figure (5.26) Show The result of proposed system 97.25% overall training of classification accuracy, it has a percentage of 98.83% sensitivity on account of high true positive , a percentage of 95.69 % specificity were present, and percentage of 2.74% error rate .

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1 Conclusion

The clinical laboratory doctor to diagnosis COVID 19 is more susceptible to infection and mistakes in taking decision and classifying is the key driver of designing CAD systems. So the proposed method was implemented to overcome this problem. This system for detecting abnormality in the lung have been successfully built and all the objectives of the project have been reached and achieved, starting with preprocessing stage by applying wiener filter, after that comes segmentation stage which consist of two phases segmentation of lung and segmentation of COVID 19, then from k mean clustering the Haralick features were extracted and selected, finally the random forests algorithm was used to classify heart sounds into normal and abnormal cases, all have been implemented using MATLAB.

The proposed system shows a result of 97.25% overall training of classification accuracy, it has a percentage of 98.83% sensitivity on account of high true positive , a percentage of 95.69 % specificity were present, and percentage of 2.74% error rate .

6.2 Recommendation

A future work on this system could include:

- Using the deep learning to classify of COVID 19.
- Using the graphical features in the classification process to increase the accuracy of the system.
- The proposed algorithm of this study can be implemented within the CT machine to supply the radiologist with real time diagnosis.

References

- 1]Worldometer, march2021, <https://www.worldometers.info/coronavirus/>.
- 2] F. Shi, J. Wang, J. Shi et al., “Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19,” IEEE Reviews in Biomedical Engineering, 2020.
- 3] Bai H., Hsieh B., Xiong Z., et al. Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT.2020
- 4] Fang Y., Zhang H., Xie J., et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR.2020
- 5] <https://www.lung.org/lung-health-diseases/lung-disease-lookup/covid-19>
- 6] Bai H. X., Hsieh B., Xiong Z., et al. Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT.2020
- 7] Fang Y., Zhang H., Xie J., Lin M. Sensitivity of chest CT for COVID19: comparison to RT-PCR. 2020
- 8] Xie X., Zhong Z., Zhao W., Zheng C., Wang F., Liu J. Chest CT for typical 2019-nCoV pneumonia: relationship to negative RT-PCR testing. 2020
- 9] Bernheim X. M., Huang M. Chest CT findings in coronavirus disease-19 (COVID19): relationship to duration of infection. 2020
- 10]Subrato Bharati, Tanvir Zaman Khan, Prajoy Podder, and Nguyen Quoc Hung. A Comparative Analysis of Image De-noising Problem : Noise Models, Denoising Filters and Applications. 2020
- 11] S.santhosh baboo and E.iyyapparaj .A classification and analysis of pulmonary nodules in CT images using random forest.2018
- 12] Sethy, P.K., Behera, S.K., Ratha, P.K.; Biswas, P. Detection of Coronavirus Disease (COVID-19) Based on Deep Features and Support Vector Machine. 2020
- 13] Prabira Kumar Sethy;Santi Kumari Behera ; Pradyumna Kumar Ratha; Preesat Biswas Detection of Coronavirus Disease (COVID-19) Based on Deep Features and Support Vector Machine.2020
- 14] Umut Özkaya, Şaban Öztürk, Serkan Budak, Farid Melgani, Kemal Polat. Classification of COVID-19 in Chest CT Images using Convolutional Support Vector Machines.2020

- 15] Umut Ozkaya, Saban Ozturk, Mucahid Barstugan. Coronavirus (COVID-19) Classification using Deep Features Fusion and Ranking Technique. 2020
- 16] Tongxue Zhou, Stéphane Canu, and Su Ruan., Automatic COVID-19 CT segmentation using U-Net integrated spatial and channel attention mechanism. 2020
- 17] Abolfazl Zargari Khuzani, Morteza Heidari, S. Ali Shariati. COVID-Classifier: an automated machine learning model to assist in the diagnosis of COVID-19 infection in chest X-ray images **2021**
- 18] Arpita Halder, and Bimal Datta. COVID-19 detection from lung CT-scan images using transfer learning approach. 2021
- 19] Arpita Halder, and Bimal Datta. COVID-19 detection from lung CT-scan images using transfer learning approach. 2021
- [20] Manikandan, Jabin Alfay, Sherin, Aadhithya, and Senthil Kumar. Classification of COVID 19 in Chest CT Images using Convolutional Neural Network. 2021
- 21] SEER Training Modules, Module Name. U. S. National Institutes of Health, National Cancer Institute. Day Month Year (of access) <<https://training.seer.cancer.gov/>>
- 22] Mohammed Ali, in Handbook of Non-Invasive Drug Delivery Systems, 2010
- 23] Rebecca A. Johnson, Helio Autran de Moraes, in Fluid, Electrolyte, and Acid-Base Disorders in Small Animal Practice (Fourth Edition), 2012
- 24] Atul Sharma, Swapnil Tiwari, and Jean Louis Marty. Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): a global pandemic and treatment strategies, 2020
- 25] David K. Meyerholz, Charles W. Frevert. Comparative Anatomy and Histology (Second Edition), 2018
- 26] Joseph Feher. Quantitative Human Physiology (Second Edition), 2017
- 27] Amir Hakim, O.S. Usmani, in Reference Module in Biomedical Sciences, 2014
- 28] U.S. Centers for Disease Control and Prevention (CDC), National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases.), Human coronavirus types. 2020,

- 29] Shereen MA, Khan S, Kazmi A, Bashir N, Siddique R. COVID-19 infection: origin, transmission, and characteristics of human coronaviruses. 2020
- 30] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020
- 31] Patel R, Babady E, Theel ES, Storch GA, Pinsky BA, George K, Smith TC, Bertuzzi S. Report from the American Society for Microbiology COVID-19 International Summit, 23 March 2020
- 32] Mason RJ. Pathogenesis of COVID-19 from a cell biology perspective. 2020
- 33] Geng, MJ., Wang, LP., Ren, X. *et al.* Risk factors for developing severe COVID-19 in China: an analysis of disease surveillance data. 2021.
- 34] McIntosh K. Coronavirus disease 2019 (COVID-19). Accessed March 31, 2020.
- 35] Chunqin Long, Huaxiang Xu, and Honglu Li. Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT. 2020
- [36] Laboratory testing for 2019 novel coronavirus (2019-nCoV) in suspected human cases". World Health Organization (*WHO*). 2020.
- 37] Li Y, Xia L. "Coronavirus Disease 2019 (COVID-19): Role of Chest CT in Diagnosis and Management". 2020
- 38] Wastnedge EA, Reynolds RM, van Boeckel SR, Stock SJ, Denison FC, Maybin JA, Critchley HO. January 2021
- 39] Carlos Cordon-Cardo, Elisabet Pujadas, and David L. Reich COVID-19: Staging of a New Disease. 2020
- 40] <https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html>
- 41] Liu S.T.H., Lin H.M., Baine I., Wajnberg A., Gumprecht J.P., Rahman F., Rodriguez D., Tandon P., Bassily-Marcus A., Bander J. Convalescent plasma treatment of severe COVID-19: a propensity score-matched control study. 2020

- 42] Jose R.J., Manuel A. COVID-19 cytokine storm: the interplay between inflammation and coagulation.2020
- 43] Nadkarni G.N., Lala A., Bagiella E., Chang H.L., Moreno P., Pujadas E., Anticoagulation, mortality, bleeding and pathology among patients hospitalized with COVID-19: a single health system study. 2020
- 44] www.fda.gov/radiation-emitting-products/medical-x-ray-imaging/computed-tomography-ct
- 45] www.ssla.co.uk/digital-image-processing/
- 46] https://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/VELDHUIZEN/node15.html
- 47] Pavan Vadapalli. Image Segmentation Techniques (Step By Step Implementation),2021
- 48] <https://www.upgrad.com/blog/image-segmentation-techniques/>
- 49] D. Gadkari, “Image quality analysis using glcm,” 2004.
- 50] D. Lu, “A Survey Of Image Classification Methods And Techniques For Improving Classification Performance”, International Journal Of Remote Sensing, 2007.
- 51] Ayubu Hassan Mbagu, " Pap Smear Images Classification For Early Detection Of Cervical Cancer", Tianjin University Of Technology And Education, May 2015.
- 52] Chaoxin Zheng ,Da-Wen Sun, and Liyun Zheng. Recent applications of image texture for evaluation of food qualities - A review,2006.
- 53]Edoras,"Image Processing Toolbox”, [Online]. Available: <https://Edoras.Sdsu.Edu/Doc/Matlab/Toolbox/Images/Regionprops.html>. [Accessed 31 June 2021].
- 54] <http://www.cyto.purdue.edu/cdroms/micro2/content/education/wirth06.pdf>. [Accessed 31 June 2021].
- 55] <https://www.investopedia.com/terms/a/autocorrelation.asp>.
- 56] <https://canvas.stanford.edu>

- 57] Tommy Löfstedt, Patrik Brynolfsson , Thomas Asklund, Tufve Nyholm, Anders Garpebring Gray-level invariant Haralick texture features. 2019
- 58] KUMAR, A., 7 Most Common Machine Learning Tasks & Related Methods.2015.
- 59] Groth, P., et al., The Semantic Web – ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings. 2016
- 60] V, V.T., et al., A Survey and Comparison of Artificial Intelligence Techniques for Image Classification and Their Applications International Journal of Science and Research 2016. 5(4): p. 188.
- 61] Polamuri, S., HOW THE RANDOM FOREST ALGORITHM WORKS IN MACHINE LEARNING. Dataaspirant, May 22, 2017
- 62] Cimss, "WhatIsMatlab"[Online].Available:
<http://Cimss.Ssec.Wisc.Edu/Wxwise/Class/Aos340/Spr00/Whatismatlab.html>.
[Accessed 2 July 2021].
- 63] C. M. S.-P. M. P. Bram Van Ginneken, "Computer-Aided Diagnosis: How To Move From The Laboratory To The Clinic",Radiology, Vol. 261, P. 3, 2011.

Appendix

Code in matlab

```
addpath('lib');
folders = {'Covid','NonCovid'};
[flst]= fulldir(fullfile('dataset',folders),'*g',true);
trgs = [flst.target];
features = [];
targets = [];
feature_pack = {};
%%
for iter = 1:numel(flst)
    %% image aquisition
    fprintf('processing image %s : %g of
%g\n',flst(iter).name,iter,numel(flst));
    img_org = imread(flst(iter).file);
    if ~ismatrix(img_org)
        img_org = rgb2gray(img_org);
    end

    %% image preprocessing
    img_org = wiener2(img_org,[3 3]);

    %% image segmentation
    lungs_mask = extract_lungs(img_org);
    if lungs_mask == 0
        continue;
    end
    km_seg = imKMseg(img_org,3);
    covid_mask = ismember(km_seg,3) & lungs_mask ;
    covid_mask = bwareafilt(covid_mask,[10 inf]);

    if isempty(find(covid_mask == 1,1))
        continue;
    end

    %% features extraction
```

```

    [ftr,trg] = lung_features(img_org,covid_mask,trgs(:,iter));
    feature_pack{iter} = ftr;
    features = [features ftr];
    targets = [targets trg];

end
save('dataset/lung_data.mat', 'features', 'targets', 'feature_pack');

%% training random forest classifier
% targets = vec2ind(targets);
mdl = trainRFC(features,targets);

%%
ptTragets = [];
actualTragets = [];
for iter = 1:numel(feature_pack)
    fprintf('evaluating: %u of %u \n',iter, numel(feature_pack))
    if ~isempty(feature_pack{iter})
        pt = predict(mdl,feature_pack{iter});
        dec = classTemplate(pt, {1, 2});
        ptTragets = [ptTragets dec.overallClass{1}];
        actualTragets = [actualTragets trgs(iter)];
    end
end
cp = classperf(actualTragets,ptTragets);
fprintf('\nCorrect Rate = %1.3f',cp.CorrectRate*100)
fprintf('\nSensitivity = %1.3f',cp.Sensitivity*100)
fprintf('\nSpecificity = %1.3f',cp.Specificity*100)
fprintf('\nError Rate = %1.3f\n',cp.ErrorRate*100)
%% saving
save('lib/lung_model.mat','mdl')

```

Table is some result of texture after the selection using Random Frost algorithm

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
21.4	0.3	0.9	0.9	214.1	-6.0	0.3	0.1	2.6	0.9
44.6	1.1	0.8	0.8	309.1	-29.6	0.7	0.1	2.6	0.7
29.3	1.2	0.8	0.8	295.3	-6.9	0.9	0.1	3.1	0.6
27.6	1.1	0.9	0.9	225.0	2.3	0.6	0.1	2.2	0.8
43.0	2.6	0.9	0.9	1314.7	-113.8	1.0	0.2	2.2	0.7
27.2	0.3	1.0	1.0	729.9	-30.5	0.3	0.1	2.6	0.8
22.9	1.2	1.0	1.0	1227.8	43.1	0.8	0.1	2.3	0.7
38.1	3.4	0.7	0.7	772.0	-63.2	1.4	0.1	2.7	0.5
21.4	3.4	0.3	0.3	121.8	-1.7	1.3	0.0	3.4	0.6
33.3	9.0	0.3	0.3	637.4	-43.5	2.3	0.1	2.2	0.5
33.2	1.3	0.9	0.9	1395.2	-24.5	0.6	0.2	2.1	0.8
19.8	0.7	0.9	0.9	262.7	-1.1	0.6	0.1	2.8	0.7
16.3	1.8	0.9	0.9	469.2	33.9	1.1	0.1	2.6	0.6
35.0	5.2	0.6	0.6	634.3	-47.3	1.9	0.2	1.9	0.5
39.1	0.4	1.0	1.0	696.6	-58.1	0.4	0.1	2.2	0.8
33.7	0.5	0.9	0.9	547.0	-21.1	0.4	0.1	2.9	0.8
31.7	1.7	0.8	0.8	416.0	-18.0	1.0	0.0	3.1	0.6
27.8	2.1	0.9	0.9	1279.6	8.8	0.9	0.1	2.6	0.7
40.9	2.0	0.7	0.7	419.2	-36.3	1.1	0.1	2.7	0.6
39.5	1.5	0.8	0.8	737.8	-62.9	0.9	0.1	2.7	0.6
39.2	0.5	1.0	1.0	651.0	-48.4	0.5	0.1	2.4	0.8
38.8	1.1	0.9	0.9	813.3	-64.2	0.6	0.1	2.7	0.7
23.5	1.3	1.0	1.0	378.2	-10.4	1.0	0.1	2.7	0.6
32.5	1.7	0.9	0.9	725.4	-19.2	1.0	0.1	2.6	0.6
41.0	2.2	0.8	0.8	769.5	-66.5	1.2	0.1	2.4	0.6
19.8	4.4	0.6	0.6	568.9	14.2	1.4	0.0	3.5	0.6
26.7	0.6	1.0	1.0	950.0	-2.9	0.5	0.1	2.8	0.8
51.4	1.4	0.6	0.6	274.7	-28.7	0.6	0.2	1.9	0.8
32.8	1.2	1.0	1.0	1632.7	-51.3	0.7	0.2	2.0	0.8
30.5	1.9	0.7	0.7	331.2	-16.7	0.9	0.1	2.9	0.7