Arab Academy for Science, Technology and Maritime Transport

College of Computing and Information Technology

Computer Science Department

B. Sc. Final Year Project

Smart (TB) Detection

Presented By:

Reda Abdelhamid Elsewidy   ***Ahmed Yosry Akrab***

Omar Ashraf Elashmony

Supervised By:

Dr. Essam Elfakharany - Dr. Nahla Belal

JULY – 2020

DECLARATION

I hereby certify that this report, which I now submit for assessment on the programme of study leading to the award of Bachelor of Science from *Arab Academy for Science, Technology and Maritime Transport ,college of computing and information technology* , is all my own work and contains no Plagiarism. By submitting this report, I agree to the following terms:

> *Any text, diagrams or other material copied from other sources (including, but not limited to, books, journals, and the internet) have been clearly acknowledged and cited followed by the reference number used; either in the text or in a footnote/endnote. The details of the used references that are listed at the end of the report are confirming to the referencing style dictated by the final year project template and are, to my knowledge, accurate and complete.*

I have read the sections on referencing and plagiarism in the final year project template. I understand that plagiarism can lead to a reduced or fail grade, in serious cases, for the Graduation Project course.

Student Name:
Registration Number:

Signed: ________________________

Date:

Student Name:
Registration Number:

Signed: ________________________

Date:

Student Name:
Registration Number:

Signed: ________________________

Date:

ABSTRACT

Tuberculosis is a disease caused by the Tuberculosis Mycobacterium. Early identification of TB decreases mortality. Early stage TB is typically diagnosed with examination of the chest by x-ray. As the digital world gradually merges with the physical world, it has become even more difficult to develop health-care services. Tuberculosis management (TB) relies on early diagnosis and treatment at the primary level of health care. Tuberculosis (TB) is classified by an infectious agent as one of the top ten causes of death. This study aimed to use deep learning ensemble learning, to predict the disease with accurate results. The suggested algorithm in deep learning approach is to use (CNN) coevolutionary neural networks algorithm, and multiple models CNN for deep learning and (CNN as feature extractor + SVM) for the ensemble learning. The results show that the proposed CNN with large dataset achieves high accuracy and accuracy of 65% for small dataset. Ensemble learning results show that Support Vector Machine (SVM) performs well among basic learning classifiers. The final result of the ensemble learning is CNN feature extractor with SVM classifier achieved an overall accuracy 90 % on two datasets. Our study in this report show that the CNN algorithm should perform well if there is large dataset for training CNN but ensample learning produces more accurate results in this study when we use SVM as our main classifier with VGG-16 pre-trained feature extractor.

KEYWORDS: Tuberculosis, Deep Learning, Ensemble Learning, Convolutional neural network (CNN), Support vector machine (SVM), Random Forest, Decision Trees, Transfer Learning, Feature extraction, chest x-ray (CXR).

TABLE OF CONTENTS

CHAPTER 1 - INTRODUCTION

CHAPTER 2 – BACKGROUND AND RELATED WORKS

CHAPTER 3 – ANALYSIS AND DESIGN

CHAPTER 4 – EXPERIMENTS AND RESULTS

CHAPTER 5 - CONCLUSION

REFERENCES

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

TB	Tuberculosis
(CNN)	Convolutional Neural Network
(MDGs)	Millennium Development Goals
(UN)	United Nations
US	United States
(WHO)	World Health Organization
(SDGs)	Sustainable Development Goals
(GPU)	General Program of Work
API	Application Programming Interface
UML	Unified Modelling Language

TB is caused by bacteria (Mycobacterium tuberculosis) that most frequently damage the lungs. tuberculosis can be curable and avoidable. When a person with TB disease coughs, speaks, or sings, the bacteria that cause TB is transmitted through the air from person to person. TB not only affects the lungs but may also infect other parts of the body, such as the bones or the spine. TB is one of the top 10 death rows causes worldwide, according to the World Health Organization. Nearby people can breathe and get sick in these bacteria. 10 million TB patients died in 2017, while 1,6 million died of TB. TB is measured using multiple ways one of them is chest by X-rays, but the patient cannot be sure of TB or any other disease. Two forms of TB disease are present: latent TB infection and TB disease. Without making you sick, TB bacteria will live inside your body. Latent TB infection is referred to. The body will battle the bacteria to keep them from developing in certain people who breathe in TB bacteria and become infected. Many with latent tuberculosis infection are not sick, have no symptoms and are unable to transmit TB bacteria to others. The person will go from having latent TB infection to being sick with TB disease if bacteria become involved in the body. That is why people with latent TB infection are also given medication to avoid TB disease. TB bacteria most commonly develop in our lungs and can cause symptoms such as: bad tuberculosis which lasts for 3 weeks or longer, chest pain, blood coughing or sputum (deep inside our lungs) and other symptoms such as weight loss, chills, fevers, sweating by night. This may lead to delays in the search for treatment and the bacteria are passed to others. Through close contacts during, persons with TB can infect up to 10-15 other persons. Without proper treatment of TB, the infected person will die. Growing prevalence of artificial intelligence (AI) and related technologies in business and society is beginning to be applied to health care. Deep learning produces hierarchical characteristics automatically from images. Research has also been performed on the use of chest x-ray images to detect TB. In the x-ray-lung images multi-level image enhancement was done. Backpropagation neural network was then used for classification of TB. simple CNN architectures were proved beneficial in terms of universality and stability. For example, Alex Net, VGGNet, Google Net, ResNet and DenseNet have demonstrated good results. Three-layers forms are included in CNNs. These layers are convolution, pooling and fully connected layer. A CNN architecture has been developed when these layers are stacked. The convolution layer defines the neuron output from which it is connected to local input regions by measuring the scalar product between its weights and the input region. The rectified linear unit (ReLu) aims to apply an "elementwise" activation function such as Tanh, sigmoid to the output of the activation produced by the previous layer. The pooling layer will perform down sampling of the given input, further reducing the number of parameters within that activation. The fully-connected layers will perform the same duties found in standard ANNs and attempt to produce class scores from the activations, to be used for classification. It is also suggested that ReLu may be used between these layers, as to enhance the performance. Unlike algorithms using hand engineered features, deep learning offers a hierarchical analysis of the image using a cascade of layers of non-linear processing units for “end-to-end” feature extraction and classification. The DL implementation for visual recognition research is often accomplished by three kind of approaches: a scratch model training, model fine tuning (also known as transfer learning (TL)), and pre trained models as feature extractors followed by supervised machine learning algorithm ex: support vector machine. Ensemble learning (EL) may benefit from improving results from these approaches. Better results are achieved than custom models by pre-trained CNNs. When a prediction is based on more than one model, it is referred to as ensemble learning. Ensemble eliminates the variance of predictions, thus providing predictions that are more accurate than any single model. ex: three models are ResNet, Inception-ResNet and DenseNet, thus the ensemble. these models were used as feature extractors and SVM was used as a classifier. Studies show that, deep learning has produced good results for TB detection. CNN show a robust result when it uses as feature extractor. However, most studies used CNN as features extractor that were automatically extracted then fed these features to the basic classifier (ensemble classifier).

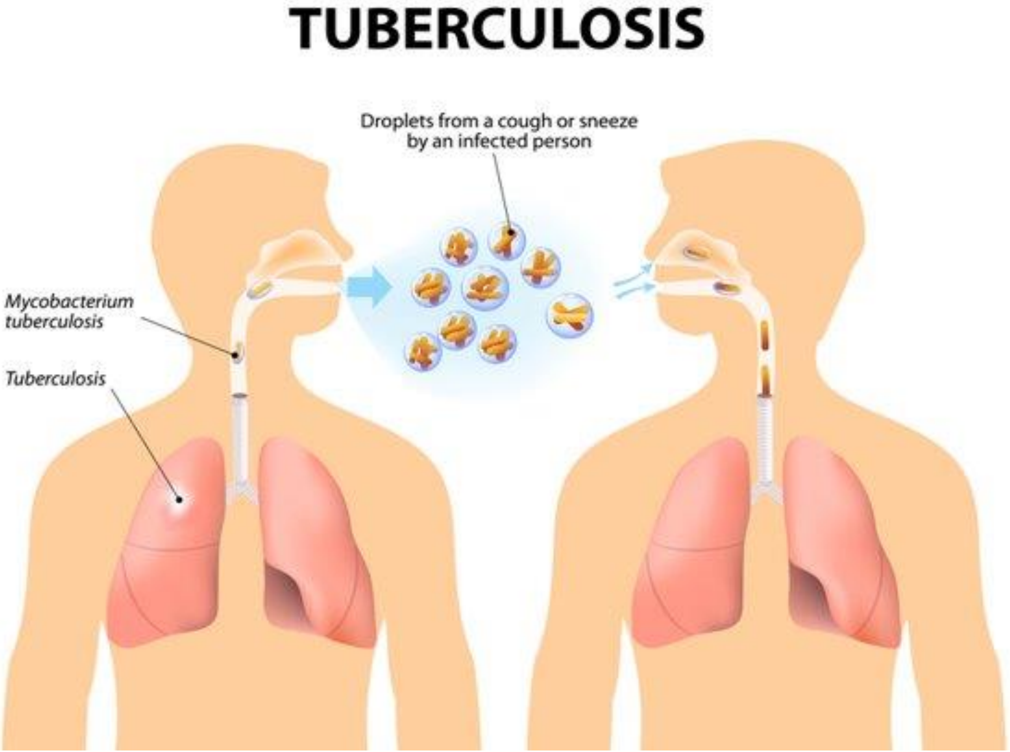

TUBERCULOSIS
Droplets from a cough or sneeze by an infected person
Mycobacterium tuberculosis
Tuberculosis

Figure 1: infection of TB

1.1 MOTIVATION

- **According to World Health Organization (WHO):**

  - In 2018, an estimated 10 million people fell ill with tuberculosis (TB) worldwide. 5.7 million men, 3.2 million women and 1.1 million children.

  - A total of 1.5 million people died from TB in 2018.

  - TB is one of the top 10 causes of death.

From 2000 to 2015, global, regional and national efforts to reduce the burden of tuberculosis (TB) disease focused on achieving targets set within the context of the Millennium Development Goals (MDGs). The MDGs were established by the United Nations (UN) in 2000, and targets were set for 2015. Target of MDG 6 was to "halt and reverse" TB incidence. The Stop TB Partnership adopted this target and set two additional targets: to halve TB prevalence and TB mortality rates by 2015 compared with their levels in 1990. The global TB strategy developed by the World Health Organization (WHO) for the decade 2006–2015 – the Stop TB Strategy – had the overall goal of reaching all three of these targets. In October 2015, WHO published its assessment of whether the 2015 global TB targets for reductions in TB incidence, prevalence and mortality had been achieved. For the period 2016–2035, global, regional and national efforts to reduce the burden of TB disease have the ambitious aim of "ending the TB epidemic", within the context of the UN's Agenda for Sustainable Development, and based on WHO's End TB Strategy. The Sustainable Development Goals (SDGs) and their associated indicators and targets were adopted by all UN Member States in September 2015. The SDGs cover the period 2016–2030, and the End TB Strategy is for the period 2016–2035. In 2017 and 2018, TB commitments included in the SDGs and End TB Strategy were reaffirmed at the first-ever global ministerial conference on TB (held in Moscow in November 2017), and the first-ever UN high-level meeting on TB (held at UN headquarters in New York in September 2018). Targets for TB that are consistent with those set in the End TB Strategy have been included in WHO's Thirteenth General Program of Work, 2019–2023 (GPW 13). According to the Sustainable Development Goals (17 SDGs) by 2030, end the tuberculosis disease.

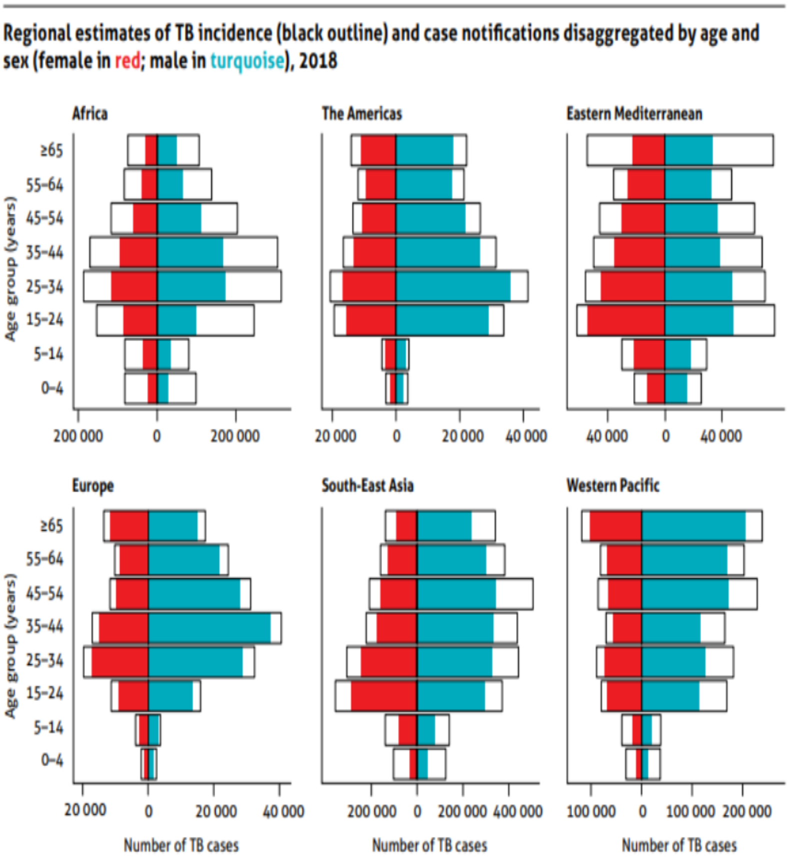
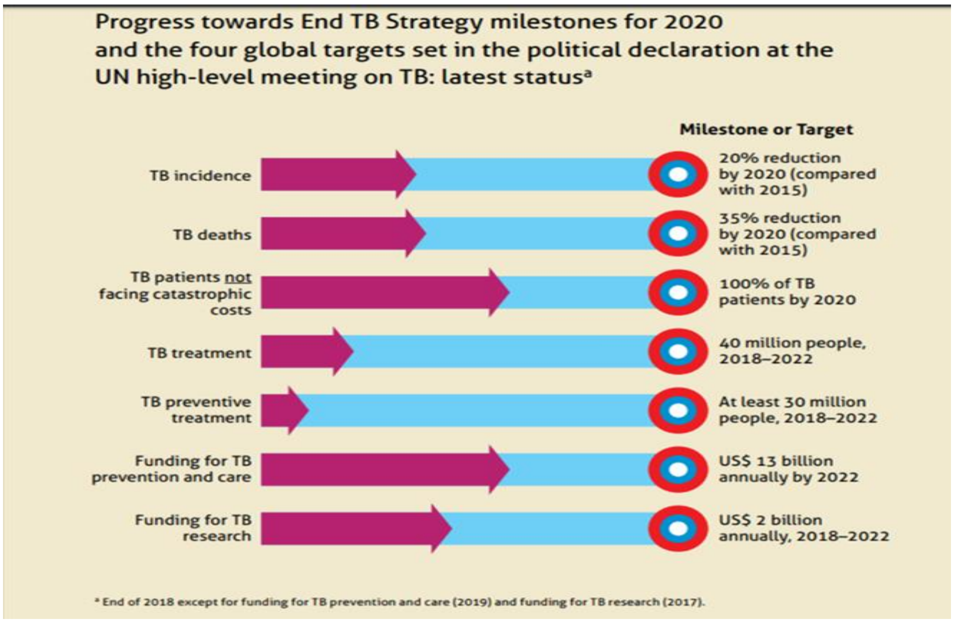

Regional estimates of TB incidence (black outline) and case notifications disaggregated by age and sex (female in red; male in turquoise), 2018

Figure 2: number of TB cases in the world


Progress towards End TB Strategy milestones for 2020 and the four global targets set in the political declaration at the UN high-level meeting on TB: latest status[a]

	Milestone or Target
TB incidence	20% reduction by 2020 (compared with 2015)
TB deaths	35% reduction by 2020 (compared with 2015)
TB patients not facing catastrophic costs	100% of TB patients by 2020
TB treatment	40 million people, 2018–2022
TB preventive treatment	At least 30 million people, 2018–2022
Funding for TB prevention and care	US$ 13 billion annually by 2022
Funding for TB research	US$ 2 billion annually, 2018–2022

[a] End of 2018 except for funding for TB prevention and care (2019) and funding for TB research (2017).

Figure 3: Milestones and cost for diagnose TB

1.2 PROBLEM STATEMENT

The available datasets are X-rays, the system should detect if a patient has TB or not from the available datasets. The main purpose for our project is to use artificial intelligence in the early detection of tuberculosis (TB) with accurate results for large datasets. The early detection is considered the most important step to recover from (TB). Another problem is making an easy method of communication between the patients and hospitals or specialized doctors. The most complex forms of machine learning involve deep learning, or neural network models with many levels of features or variables and hidden layers that predict outcomes especially CNN algorithm that we will use as feature extractor in our work. After that we will build the system portal that will make each user is able to act on his functions through it. We aim to make the specialized doctors able to diagnose the X-ray test to know whether the (TB) is positive or negative by using our system.

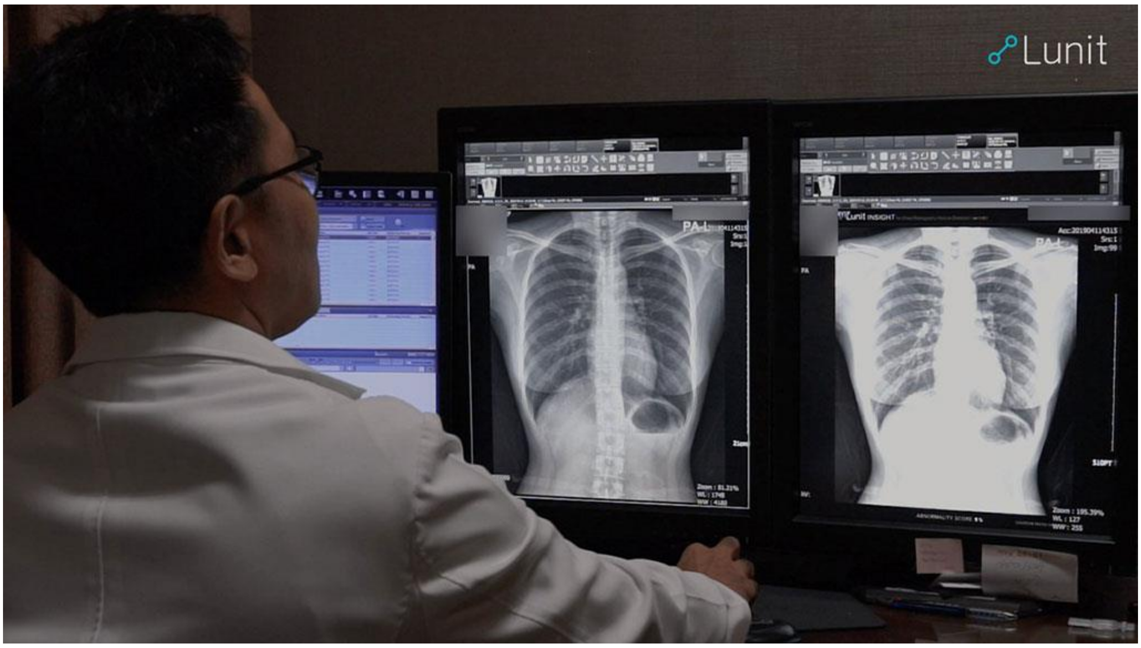
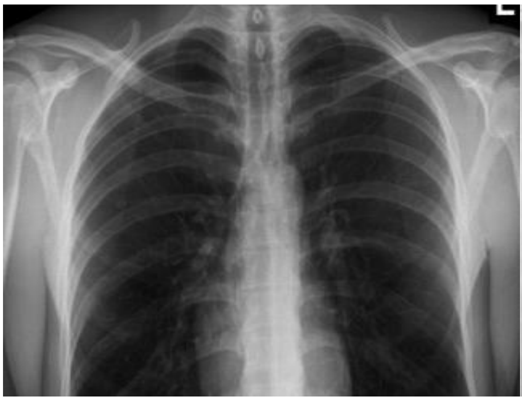


Figure 4: show how doctors use Smart (TB)



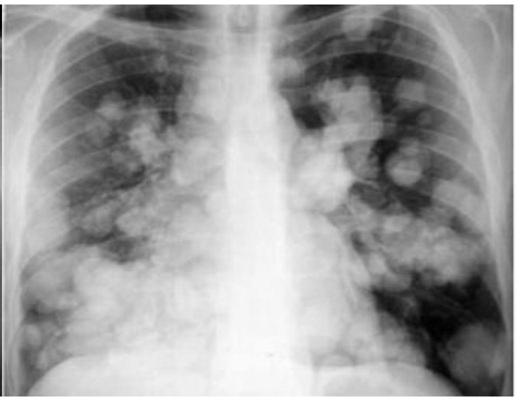


Figure 5: chest x-ray without TB **Figure 6: chest x-ray with TB**

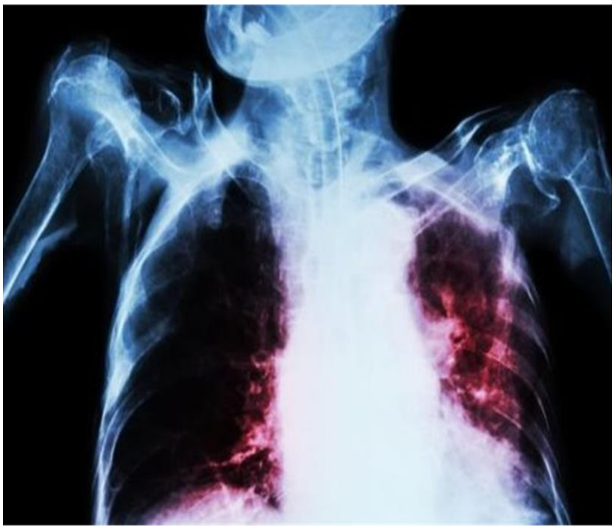


Figure 7: Severe lungs atrophy and internal bleeding

1.3 OBJECTIVES

Our objective is to build the model that takes X-ray images using Ensemble Learning model. Then we will build the Smart (TB) portal. This portal will be user friendly. we aim to develop an interface which any user can act on it without having an IT background. By using Smart (TB) portal each specialized doctor will be able to use this portal to diagnose the X-ray test to know whether the (TB) is positive or negative. Also, patients will be able to communicate easily with the specialized doctors and hospitals to book for the X-ray test and know from them the early diagnosis for their X-ray tests through the Smart (TB) portal.

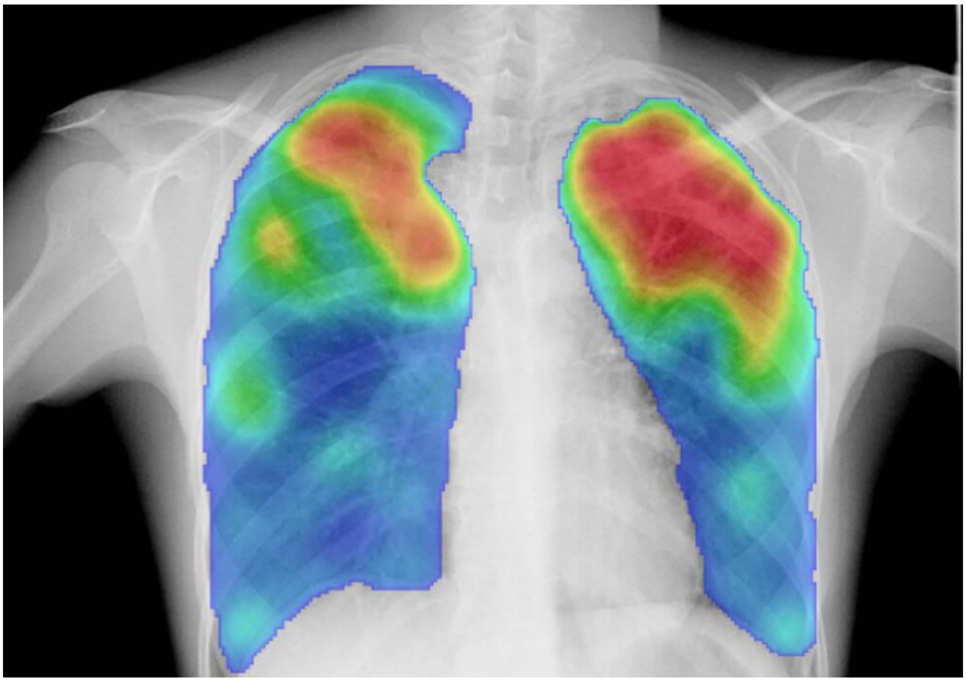
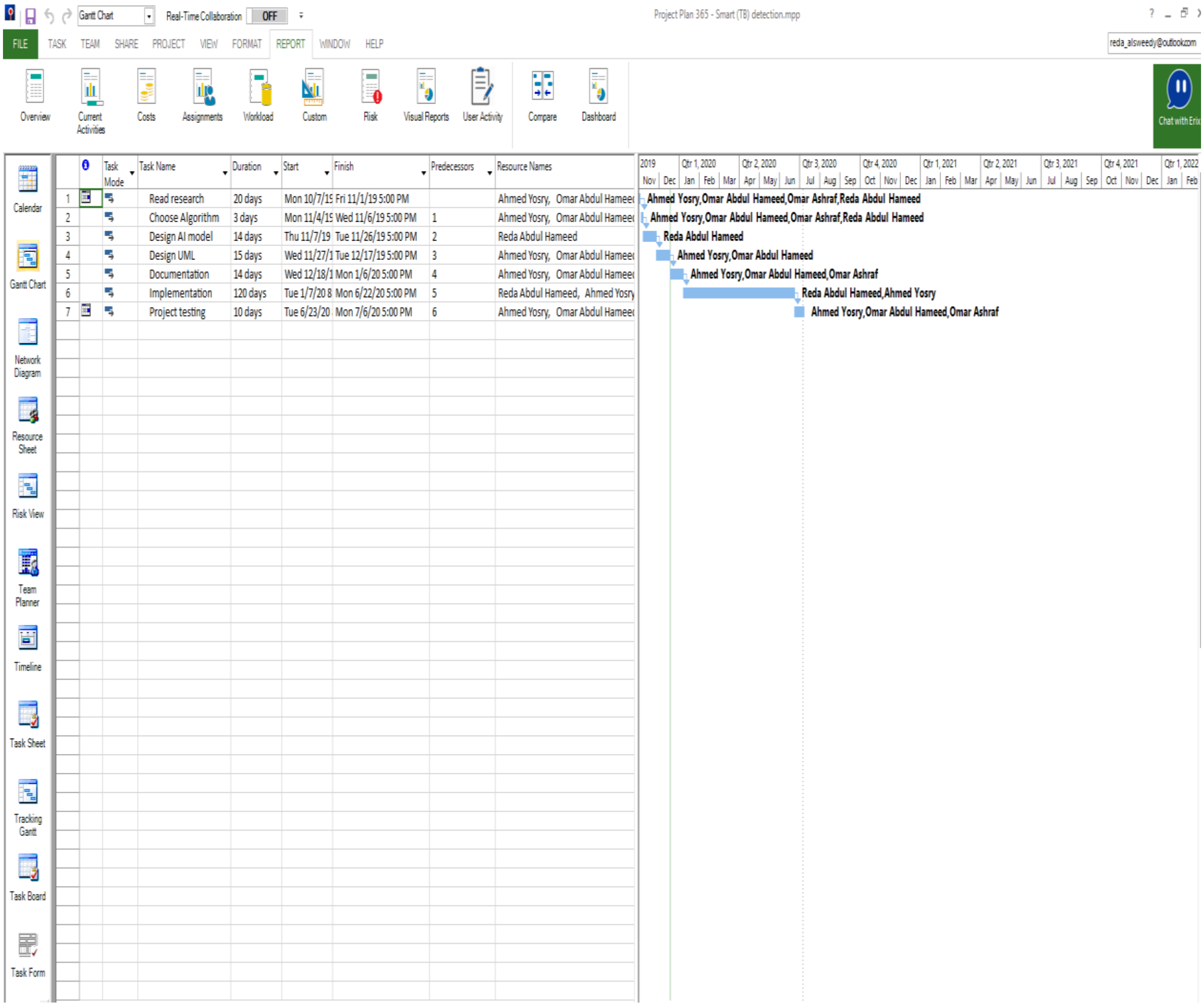


Figure 8: Thermal image shows the position of infection of TB

1.4 TIME PLAN MANAGEMENT FOR SMART (TB)

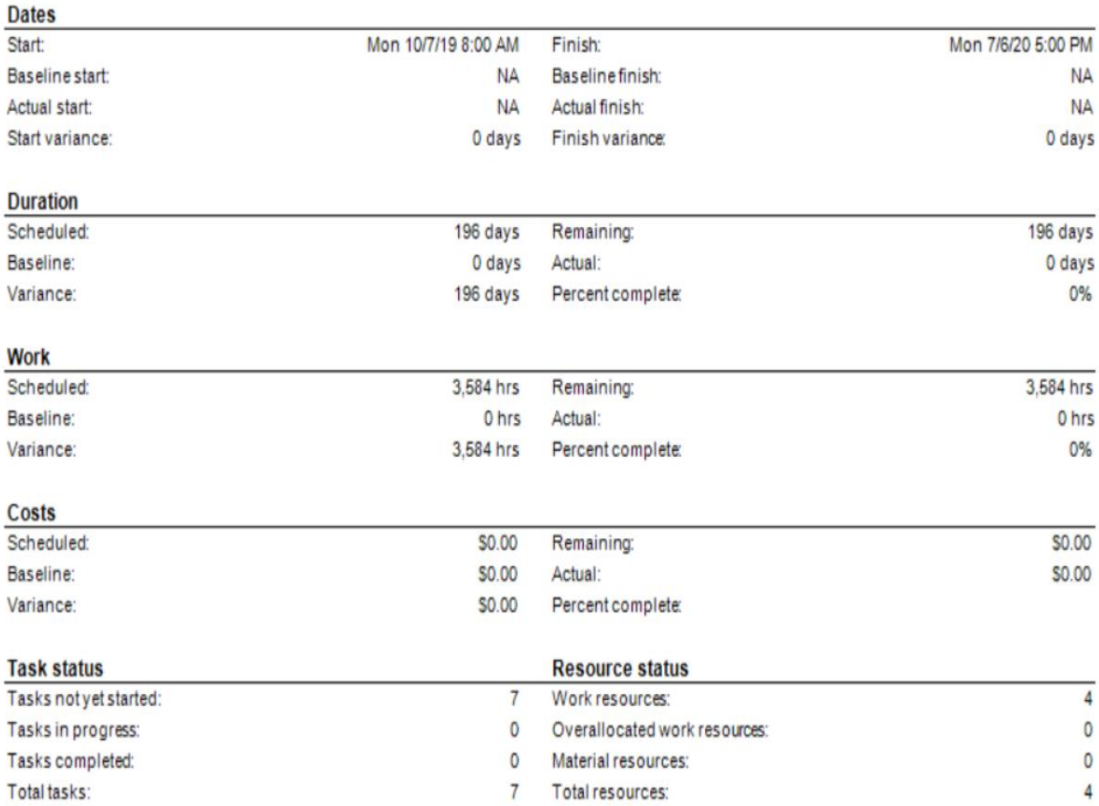


	ℹ	Task Mode	Task Name	Duration	Start	Finish	Predecessors	Resource Names
1			Read research	20 days	Mon 10/7/19	Fri 11/1/19 5:00 PM		Ahmed Yosry, Omar Abdul Hamee
2			Choose Algorithm	3 days	Mon 11/4/19	Wed 11/6/19 5:00 PM	1	Ahmed Yosry, Omar Abdul Hamee
3			Design AI model	14 days	Thu 11/7/19	Tue 11/26/19 5:00 PM	2	Reda Abdul Hameed
4			Design UML	15 days	Wed 11/27/1	Tue 12/17/19 5:00 PM	3	Ahmed Yosry, Omar Abdul Hamee
5			Documentation	14 days	Wed 12/18/1	Mon 1/6/20 5:00 PM	4	Ahmed Yosry, Omar Abdul Hamee
6			Implementation	120 days	Tue 1/7/20 8	Mon 6/22/20 5:00 PM	5	Reda Abdul Hameed, Ahmed Yosry
7			Project testing	10 days	Tue 6/23/20	Mon 7/6/20 5:00 PM	6	Ahmed Yosry, Omar Abdul Hamee

Figure 9: Time plan management for Smart (TB)

Our project plan is powered by project plan 365 project management tool.

Project summary:



Dates			
Start:	Mon 10/7/19 8:00 AM	Finish:	Mon 7/6/20 5:00 PM
Baseline start:	NA	Baseline finish:	NA
Actual start:	NA	Actual finish:	NA
Start variance:	0 days	Finish variance:	0 days

Duration			
Scheduled:	196 days	Remaining:	196 days
Baseline:	0 days	Actual:	0 days
Variance:	196 days	Percent complete:	0%

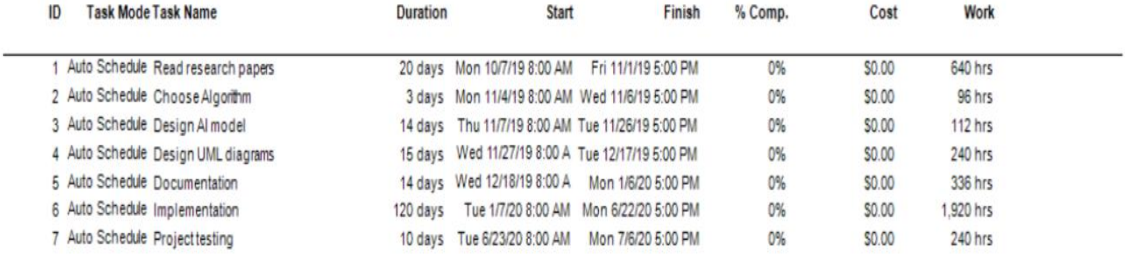
Work			
Scheduled:	3,584 hrs	Remaining:	3,584 hrs
Baseline:	0 hrs	Actual:	0 hrs
Variance:	3,584 hrs	Percent complete:	0%

Costs			
Scheduled:	$0.00	Remaining:	$0.00
Baseline:	$0.00	Actual:	$0.00
Variance:	$0.00	Percent complete:	

Task status		Resource status	
Tasks not yet started:	7	Work resources:	4
Tasks in progress:	0	Overallocated work resources:	0
Tasks completed:	0	Material resources:	0
Total tasks:	7	Total resources:	4

Figure 10: Project summary

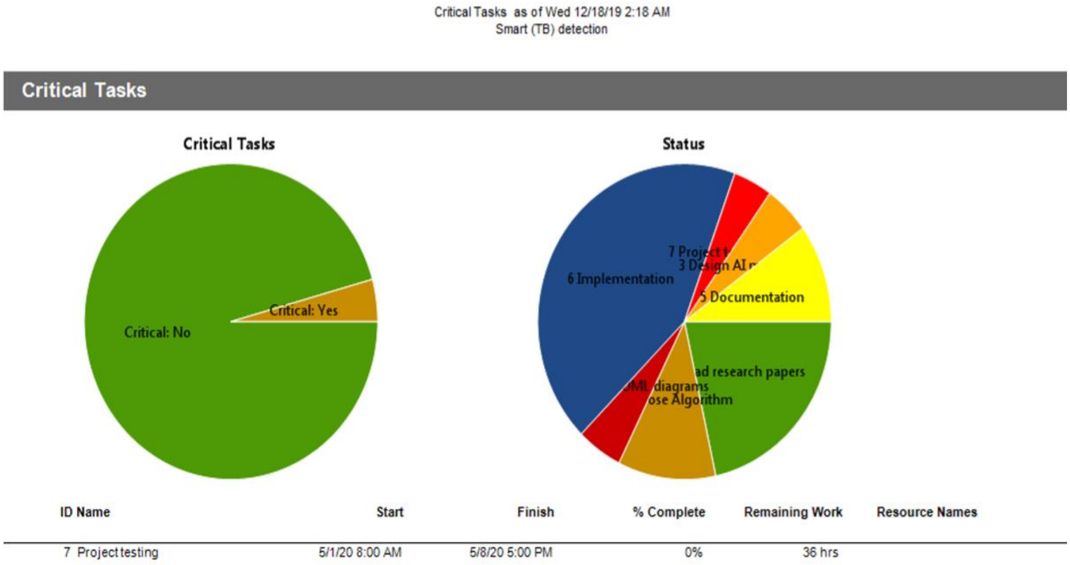
Top level tasks:



ID	Task Mode	Task Name	Duration	Start	Finish	% Comp.	Cost	Work
1	Auto Schedule	Read research papers	20 days	Mon 10/7/19 8:00 AM	Fri 11/1/19 5:00 PM	0%	$0.00	640 hrs
2	Auto Schedule	Choose Algorithm	3 days	Mon 11/4/19 8:00 AM	Wed 11/6/19 5:00 PM	0%	$0.00	96 hrs
3	Auto Schedule	Design AI model	14 days	Thu 11/7/19 8:00 AM	Tue 11/26/19 5:00 PM	0%	$0.00	112 hrs
4	Auto Schedule	Design UML diagrams	15 days	Wed 11/27/19 8:00 A	Tue 12/17/19 5:00 PM	0%	$0.00	240 hrs
5	Auto Schedule	Documentation	14 days	Wed 12/18/19 8:00 A	Mon 1/6/20 5:00 PM	0%	$0.00	336 hrs
6	Auto Schedule	Implementation	120 days	Tue 1/7/20 8:00 AM	Mon 6/22/20 5:00 PM	0%	$0.00	1,920 hrs
7	Auto Schedule	Project testing	10 days	Tue 6/23/20 8:00 AM	Mon 7/6/20 5:00 PM	0%	$0.00	240 hrs

Figure 11: Top level tasks

Critical tasks:



Critical Tasks as of Wed 12/18/19 2:18 AM
Smart (TB) detection

Critical Tasks

ID Name	Start	Finish	% Complete	Remaining Work	Resource Names
7 Project testing	5/1/20 8:00 AM	5/8/20 5:00 PM	0%	36 hrs	

Figure 12: Critical tasks

1.5 ORGANIZATION OF REPORT

Chapter 1 – Introduction: It gives information about Tuberculosis(TB) disease nature, also this chapter describes our motivation, problem statement and objectives.

Chapter 2 – Background and related works: In this chapter we talk about related works to our project and what is the similar systems to our system in the manner of purpose and problem that these systems solve. Also in this chapter we described our AI model and its background.

Chapter 3 – Analysis and design: This chapter describes our system functions, quality attributes(Non-functional requirements), how our business runs as a software product, our system design , what are the methods we used for our AI model and what is our plan to control risks that may affect or harm system.

Chapter 4 – Experiments and results: This chapter describes briefly our experiment to work and its results.

Chapter 5 - Conclusion: This chapter is concerned to a discussion about our conclusion for our project and our future work. Also we mentioned in it our references for our research.

What is Smart (TB)? It is an AI model (Ensemble Learning) that uses chest X-rays of patients who have some/all symptoms of TB disease. A specific application (CNN algorithm) is used for extract features from chest X-rays then fed the feature vector to support vector machine (SVM). In our solution, before sending the data to the CNN model we need to enhance these images by applying different needed techniques using image processing.

BACKGROUND ACQUIRED IN TB DISEASE:

Tuberculosis (TB) is a contagious disease caused by Mycobacterium tuberculosis. Like the common cold, it spreads through the air. People who are ill with pulmonary TB (TB of the lungs, the site most commonly affected) are often infectious and can spread the disease by coughing, sneezing or simply talking, as these acts propel TB bacteria into the air. Another person breathing in the bacteria may become infected with TB but will not necessarily become sick with the disease. In this case, the TB skin test will show positive. If the bacteria go on to overcome the body’s immune system, the person then becomes ill with TB. A person ill with TB presents different symptoms depending on the site of the body affected. In pulmonary TB, common symptoms are a cough with sputum production (sometimes with blood), shortness of breath and chest pain. There are also general symptoms such as fever in the evening, night sweats, loss of weight, loss of appetite, fatigue and muscle weakness. The main tools for the diagnosis of TB are clinical assessment, and bacteriological and radiological investigation. The examination of a sputum smear by microscopy is the simplest, cheapest and most direct way to identify the presence of TB bacteria and confirm pulmonary TB disease in one to two days. However, to evaluate drug susceptibility, the bacteria need to be cultivated and tested in a suitable laboratory for between 6 and 16 weeks. This makes it possible to identify the drug-resistant forms of TB. X-ray findings may be indicative of TB but usually need confirmation by means of other tests. The World Health Organization (WHO) estimates that one third of the world’s population is infected with TB and that 8.8 million new TB cases and 1.6 million deaths from TB occurred in the world in 2005. Eighty percent of all cases were in 22 countries, mainly in Africa and Asia. Seen from a global perspective, the WHO European Region accounts for only 5% of all TB cases and has lower incidence, prevalence and mortality than the regions mentioned above.

However, some countries in the European Region have TB incidence rates comparable to those in Africa, and the Region's overall treatment success rate is the same as that in Africa.

BACKGROUND ACQUIRED THROUGH THE TECHNOLOGIES:

Convolution neural networks (CNNs) are like traditional ANNs, consisting of neurons which optimize themselves by learning. Every neuron still receives an x-ray as input and performs an operation (e.g. a scalar product followed by a nonlinear function)-the basis of countless ANNs. The final layer contains class-related loss functions, and all standard tips and tricks developed for traditional ANNs remain in effect. CNN consist of many layers after the input layer, convolution layer followed by activation function, pooling layer, fully connected layer followed by activation function and the last layer is output layer and we will discuss how CNN layer working in details. VGG-16 is a convolutional neural network that is 16 layers deep. VGG-16 can load a pretrained version of the network trained on more than a million images from the ImageNet database. The pretrained network can classify images into 1000 object categories. The network has an image input size of 224-by-224 rgb channel. Support vector machines so called as SVM is a *supervised learning algorithm* which can be used for classification and regression problems as support vector classification (SVC) and support vector regression (SVR). Ensemble Learning When a classification is made by more than one model. Ensemble reduces the variance in the predictions and therefore provides more accurate predictions than a single model.

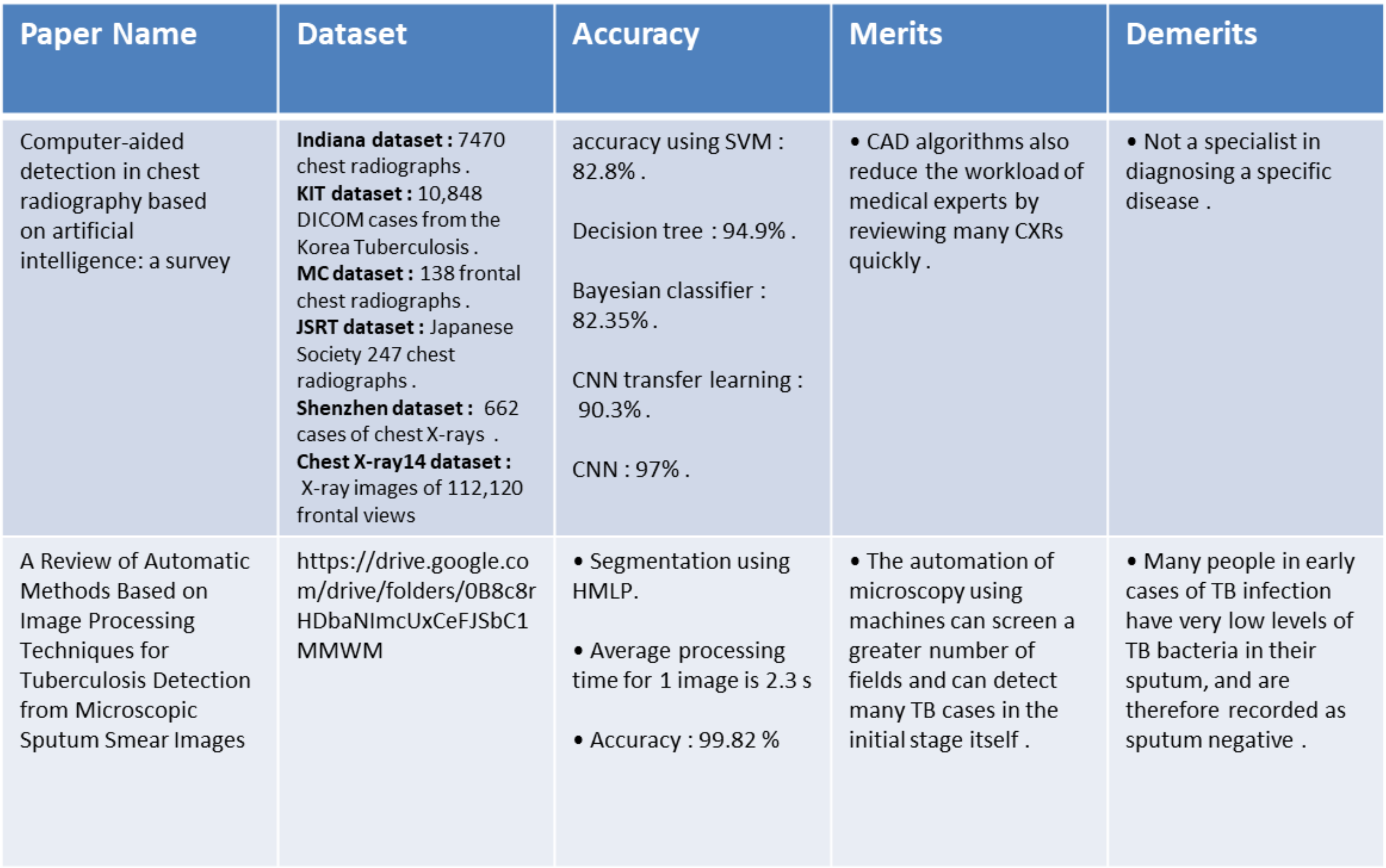
Convolutional Layer Backpropagation in ANN:

$$\mathbf{dA}^{[l]} = \frac{\partial L}{\partial \mathbf{A}^{[l]}} \quad \mathbf{dZ}^{[l]} = \frac{\partial L}{\partial \mathbf{Z}^{[l]}} \quad \mathbf{dW}^{[l]} = \frac{\partial L}{\partial \mathbf{W}^{[l]}} \quad \mathbf{db}^{[l]} = \frac{\partial L}{\partial \mathbf{b}^{[l]}}$$

$$\mathbf{dZ}^{[l]} = \mathbf{dA}^{[l]} * g'(\mathbf{Z}^{[l]})$$

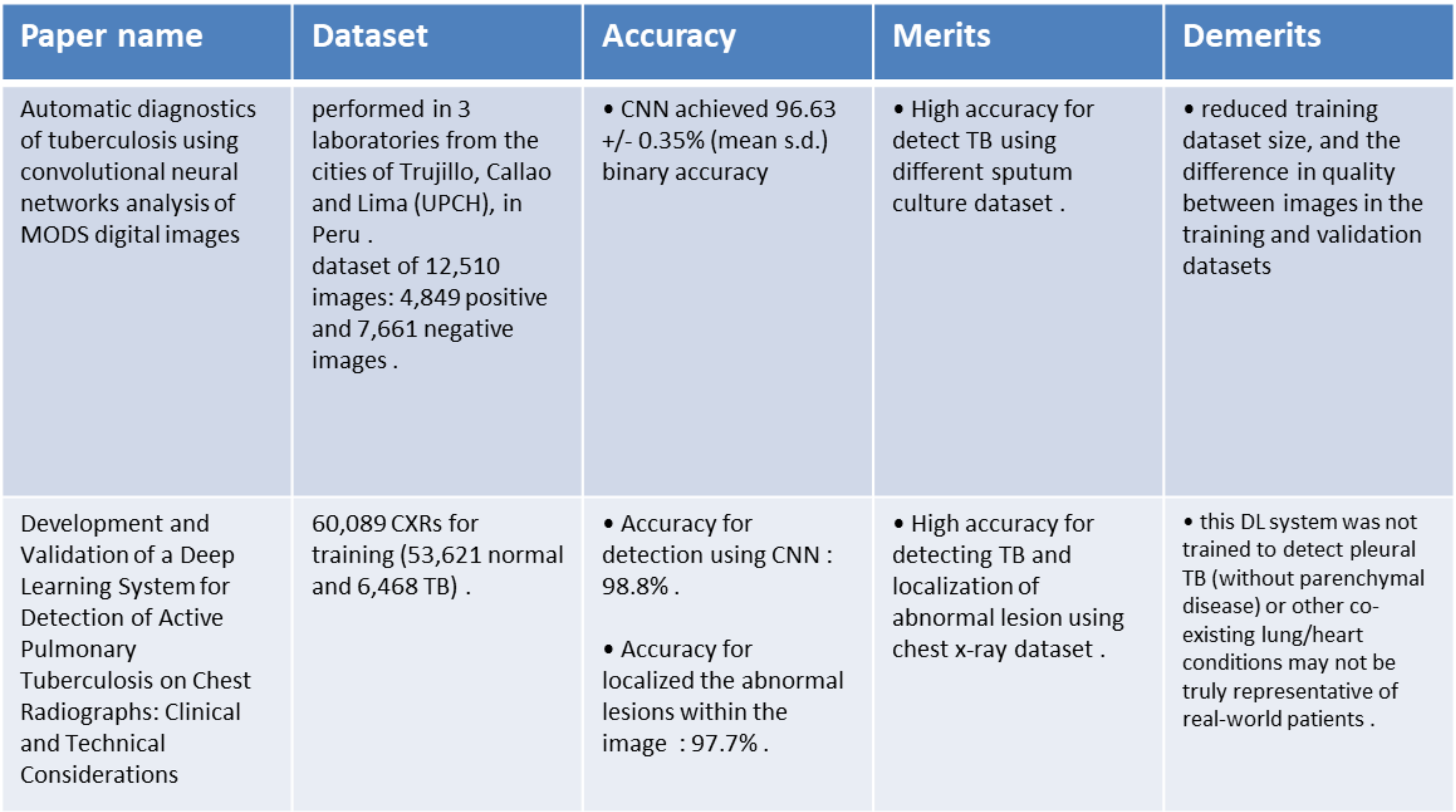
$$\mathbf{dA} += \sum_{m=0}^{n_h} \sum_{n=0}^{n_w} \mathbf{W} \cdot \mathbf{dZ}[m,n]$$

2.1 RELATED WORKS



Paper Name	Dataset	Accuracy	Merits	Demerits
Computer-aided detection in chest radiography based on artificial intelligence: a survey	**Indiana dataset :** 7470 chest radiographs . **KIT dataset :** 10,848 DICOM cases from the Korea Tuberculosis . **MC dataset :** 138 frontal chest radiographs . **JSRT dataset :** Japanese Society 247 chest radiographs . **Shenzhen dataset :** 662 cases of chest X-rays . **Chest X-ray14 dataset :** X-ray images of 112,120 frontal views	accuracy using SVM : 82.8% . Decision tree : 94.9% . Bayesian classifier : 82.35% . CNN transfer learning : 90.3% . CNN : 97% .	• CAD algorithms also reduce the workload of medical experts by reviewing many CXRs quickly .	• Not a specialist in diagnosing a specific disease .
A Review of Automatic Methods Based on Image Processing Techniques for Tuberculosis Detection from Microscopic Sputum Smear Images	https://drive.google.com/drive/folders/0B8c8rHDbaNImcUxCeFJSbC1MMWM	• Segmentation using HMLP. • Average processing time for 1 image is 2.3 s • Accuracy : 99.82 %	• The automation of microscopy using machines can screen a greater number of fields and can detect many TB cases in the initial stage itself .	• Many people in early cases of TB infection have very low levels of TB bacteria in their sputum, and are therefore recorded as sputum negative .

Table 1: some of related work on TB disease

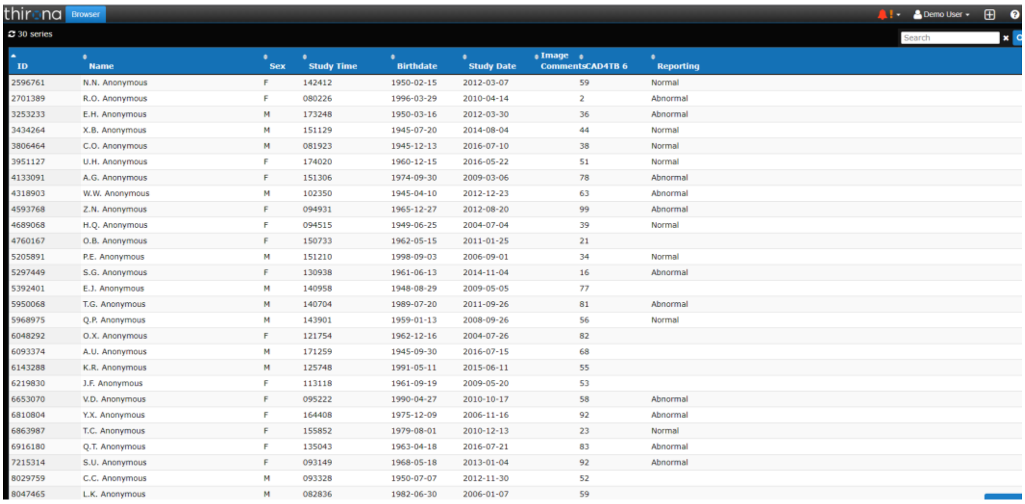


Paper name	Dataset	Accuracy	Merits	Demerits
Automatic diagnostics of tuberculosis using convolutional neural networks analysis of MODS digital images	performed in 3 laboratories from the cities of Trujillo, Callao and Lima (UPCH), in Peru . dataset of 12,510 images: 4,849 positive and 7,661 negative images .	• CNN achieved 96.63 +/- 0.35% (mean s.d.) binary accuracy	• High accuracy for detect TB using different sputum culture dataset .	• reduced training dataset size, and the difference in quality between images in the training and validation datasets
Development and Validation of a Deep Learning System for Detection of Active Pulmonary Tuberculosis on Chest Radiographs: Clinical and Technical Considerations	60,089 CXRs for training (53,621 normal and 6,468 TB) .	• Accuracy for detection using CNN : 98.8% . • Accuracy for localized the abnormal lesions within the image : 97.7% .	• High accuracy for detecting TB and localization of abnormal lesion using chest x-ray dataset .	• this DL system was not trained to detect pleural TB (without parenchymal disease) or other co-existing lung/heart conditions may not be truly representative of real-world patients .

Table 2: some of related work on TB disease

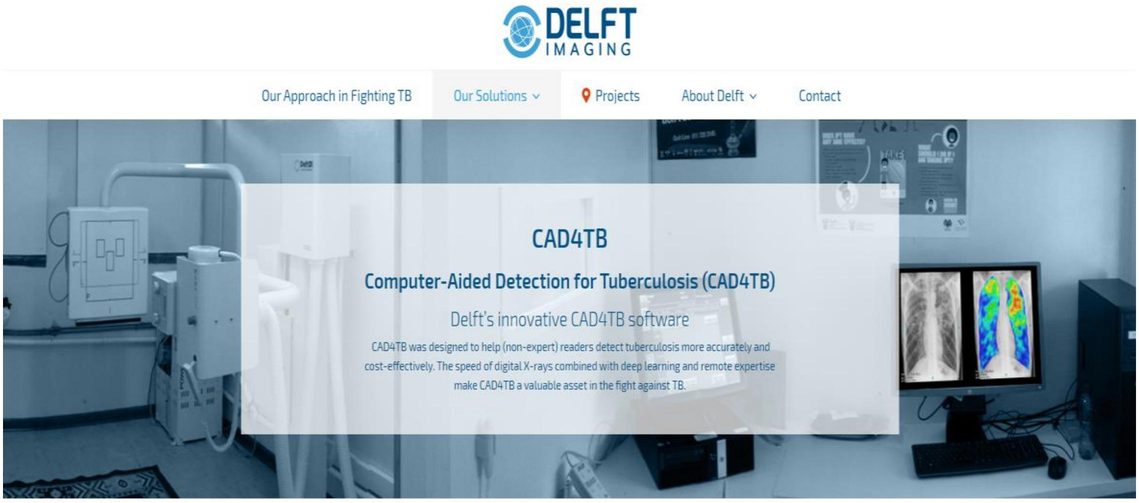
2.1.1 SIMILAR SYSTEM

Computer -aided detection for tuberculosis (CAD4TB): CAD4TB was designed to help (non-expert) readers detect tuberculosis more accurately and cost-effectively. using deep learning algorithm with digital x-rays for detection.



ID	Name	Sex	Study Time	Birthdate	Study Date	Image Comments	CAD4TB 6	Reporting
2596761	N.N. Anonymous	F	142412	1950-02-15	2012-03-07		59	Normal
2701389	R.O. Anonymous	F	080226	1996-03-29	2010-04-14		2	Abnormal
3253233	E.H. Anonymous	M	173248	1950-03-16	2012-03-30		36	Abnormal
3434264	X.B. Anonymous	M	151129	1945-07-20	2014-08-04		44	Normal
3806464	C.O. Anonymous	M	081923	1945-12-13	2016-07-10		38	Normal
3951127	U.H. Anonymous	F	174020	1960-12-15	2016-05-22		51	Normal
4133091	A.G. Anonymous	F	151306	1974-09-30	2009-03-06		78	Abnormal
4318903	W.W. Anonymous	M	102350	1945-04-10	2012-12-23		63	Abnormal
4593768	Z.N. Anonymous	F	094931	1965-12-27	2012-08-20		99	Abnormal
4689068	H.Q. Anonymous	F	094515	1949-06-25	2004-07-04		39	Normal
4760167	O.B. Anonymous	F	150733	1962-05-15	2011-01-25		21	
5205891	P.E. Anonymous	M	151210	1998-09-03	2006-09-01		34	Normal
5297449	S.G. Anonymous	F	130938	1961-06-13	2014-11-04		16	Abnormal
5392401	E.J. Anonymous	M	140958	1948-08-29	2009-05-05		77	
5950068	T.G. Anonymous	M	140704	1989-07-20	2011-09-26		81	Abnormal
5968975	Q.P. Anonymous	M	143901	1959-01-13	2008-09-26		56	Normal
6048292	O.X. Anonymous	F	121754	1962-12-16	2004-07-26		82	
6093374	A.U. Anonymous	M	171259	1945-09-30	2016-07-15		68	
6143288	K.R. Anonymous	M	125748	1991-05-11	2015-06-11		55	
6219830	J.F. Anonymous	F	113118	1961-09-19	2009-05-20		53	
6653070	V.D. Anonymous	F	095222	1990-04-27	2010-10-17		58	Abnormal
6810804	Y.X. Anonymous	F	164408	1975-12-09	2006-11-16		92	Abnormal
6863987	T.C. Anonymous	F	155852	1979-08-01	2010-12-13		23	Normal
6916180	Q.T. Anonymous	F	135043	1963-04-18	2016-07-21		83	Abnormal
7215314	S.U. Anonymous	F	093149	1968-05-18	2013-01-04		92	Abnormal
8029759	C.C. Anonymous	M	093328	1950-07-07	2012-11-30		52	
8047465	L.K. Anonymous	M	082836	1982-06-30	2006-01-07		59	

Figure 13: CAD4TB system



CAD4TB

Computer-Aided Detection for Tuberculosis (CAD4TB)

Delft's innovative CAD4TB software

CAD4TB was designed to help (non-expert) readers detect tuberculosis more accurately and cost-effectively. The speed of digital X-rays combined with deep learning and remote expertise make CAD4TB a valuable asset in the fight against TB.

Figure 14: CAD4TB interface

2.2 COMPARATIVE STUDY FOR DIFFERENT MEDICAL TB TEST

TB test	Task	Time	Accuracy
Chest X-Ray	Preparation, positioning, processing of films and repeating any images if necessary	20 to 30 min for the procedure and preparing the X-Ray film +Time needed for interpreting the film by doctors.	The results from a chest X-ray cannot confirm that a person has TB disease. Thus, chest X-ray has poor specificity
Mantoux Tuberculin Skin Test (TST)	Inject measured amount of TB antigen called Purified Protein Derivative (PPD) under the top layer of the skin in the forearm and reading the induration within 2 days	48 to 72 h.	This is highly sensitive test, but it cannot tell us whether TB is active or not. In endemic areas like India, around 30 % of population may harbour TB bacilli and TST could turn positive in them as well as those who were given BCG vaccine. So, its use for disease screening is limited.
IGRA (Interferon	Whole blood samples	24 h.	BCG vaccination will

Gamma Release Assays) – blood tests	are taken Mixed with TB antigens and controls to measure Interferon G concentration		not affect the result. However here also one cannot differentiate between active infection and latent infection Has limited use in public health
Sputum smear microscopy	Three samples of coughed up or expectorated sputum is prepared and stained with Ziehl Neelsen stain and examined under microscope for the presence of Acid-Fast Bacilli	Within 30 min Often require three different samples collected over three days	Widely used method in public health programs although the sensitivity only about 50–60 %. Many people with HIV and TB co-infection and in early cases of TB infection have very low levels of TB bacteria in their sputum, and are therefore recorded as sputum negative.
M. Tuberculosis culture	Inoculating clinical samples in media like Lowenstein-Jensen (LJ) media and incubating at 37 C for 2 – 6 weeks.	Average of 4 weeks to get the result.	The most accurate among available tests, but of limited use due to long waiting period and need for facilities for culture
GeneXpert –	Unprocessed sputum	Less than 2 h.	It is reliable, does not

MTB/RIF	samples can be used. It is molecular test based on Nucleic Acid Amplification Method (NAAM) and looks for DNA sequences specific to M. tuberculosis and Rifampicin resistance by Polymerase Chain Reaction (PCR).		need much expertise and have the added advantage of detecting drug resistance. However, the high cost of machine limits its wider use in public health interventions.

Table 3: comparative study between different TB test

2.3 MACHINE LEARNING

2.3.1 Deep learning approach

We illustrate our deep learning approach according to our system implementation map

2.3.1.1 Convolutional neural network (CNN) as a classifier

CNN is widely use in medical approach especially in tuberculosis classification problems. When using convolutional layers followed by pooling and finally with the fully connected layers, the spatial information in the x-ray can be much better utilized. Current work is using CNN model as custom model or transfer learning. Custom model is to build a CNN architecture, weights and layers from the scratch. Transfer learning is to use a pre-trained model that is applicable for tuberculosis classification, ex: pre-trained vgg-16 image net, AlexNet and ResNet archives high accuracy for diagnose tuberculosis. This study proves that CNN is very powerful method but it requires large amounts of memory and computation, both for training and testing, and large number of degrees of freedom and data to reduce overfitting and generalize the model. CNN archives AUC of 97.7% when using dataset consist of 60,089 images in training phase but archives 65% when use dataset of 800 images. other work is to make pre-processing method yielding the best results, is a combination of taking the ROI image of only the lung region and combining this with enhancing in the contrast of the image. The process is executed three times and the result is accuracy of 91.04%.

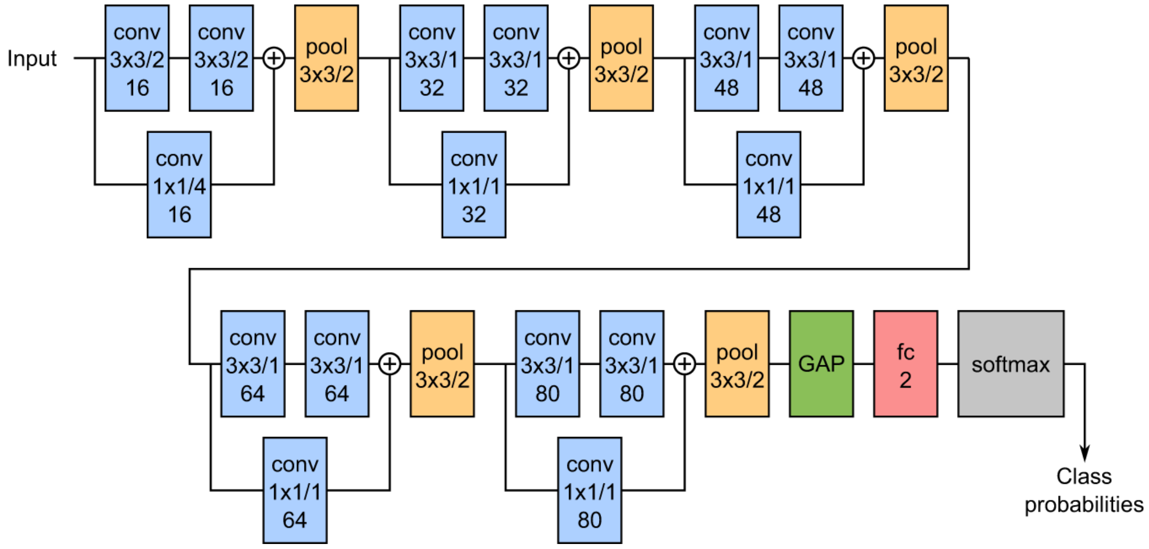


Figure 15: Example of CNN architecture

2.3.1.2 Convolutional neural network (CNN) as a feature extractor

Deep learning models have particularly revealed their autonomy in extracting useful features in the task of image classification, especially Convolution neural networks (CNNs). This feature extraction process is done by transfer learning models where pre-trained CNN models learn the generic features of large data sets such as ImageNet which are later transferred to a task. Availability in process of important function extraction of pretrained CNN models as AlexNet, VGGNet, Xception, ResNet, inception v3 and DenseNet. The classification used with highly rich derived features often demonstrates better efficiency in classification.

2.3.2 Ensemble learning approach

When a classification is made by more than one model, it is referred to as ensemble learning. Ensemble reduces the variance in the predictions and therefore provides more accurate predictions than a single model. A collection was also used for classification of TB via the feature-level merger of three deep neural network or machine learning models. Ex: by using three models ResNet, Inception-ResNet and DenseNet, and thus the ensemble has been renamed RID. The models were employed as extractors and the SVM classification was used as basic classifier. further ensemble of three architectures, AlexNet, GoogleNet and ResNet, was used for classification TB. Each architecture was built and trained from scratch, and choose different optimal hyper-parameter values. The accuracy, specificity and sensitivity of the ensemble model are higher than when used each of the standard architecture individually. The TB Classification was used with fine-tuned AlexNet, VGG-16, VGG-19, ResNet-50, ResNet-101 and ResNet-512. A group of six CNNs was created. The models of the ensemble were obtained by simply a linear mean of the predictions of probability given by each model. Pre-trained AlexNet and GoogleNet were used for the classification of pulmonary TB and they found that the pre-trained model achieved high accuracy. These models were later combined using the weighted averages of the probability values of each model and fed to SVM for the classification of Tuberculosis. Other way is to use a machine learning technique rather than using pre-trained CNN models in deep learning ex: using Support vector machine with random forest or using decision tress with bagging or boosting algorithm, in these models were able to obtain a high accuracy for CXR tuberculosis classification problem. Finally, this study shows how the importance of ensemble learning techniques are even if we use deep leaning or machine learning as the ensemble model.

the equation of the hyperplane in the 'M' dimension:

$$\begin{aligned} y &= w_0 + w_1x_1 + w_2x_2 + w_3x_3 \dots . \\ &= w_0 + \sum_{i=1}^{m} w_ix_i \\ &= w_0 + w^TX \\ &= b + w^TX \end{aligned}$$

Loss Function Interpretation of SVM:-

$$Z_i = y_i(w^Tx_i + b)$$

$$Z_i \geq 1$$

Polynomial kernel:

$$K(X_1, X_2) = (a + X_1^TX_2)^b$$

Radial basis function kernel (RBF)/ Gaussian Kernel:

$$K(X_1, X_2) = exponent(-\gamma \|X_1 - X_2\|^2)$$

the variance-covariance matrix:

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n} (Xi - \bar{x})(Yi - \bar{y})$$

Equation to Compute Eigenvectors and corresponding Eigenvalues:

$$\det(A - \lambda I) = 0$$

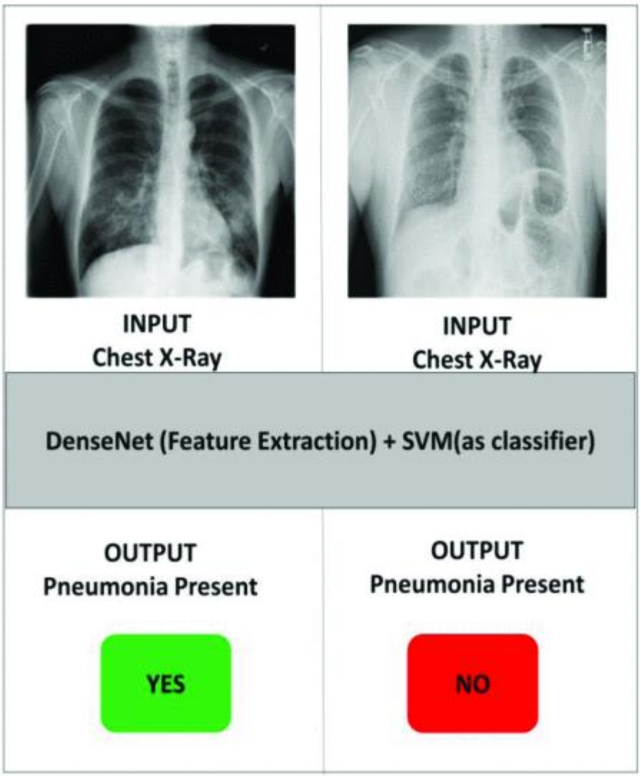


Figure 16: Example of ensemble learning with CNN DenseNet and SVM

3.1 FUNCTIONAL REQUIREMENTS FOR SMART (TB)

3.1.1 PATIENTS REQUIREMENTS

Register: Register as a patient.

Login: Login as a patient.

Book x-ray: patient can book an x-ray by specifying the patient’s location and finding the nearest laboratory surrounding it.

Book a doctor: patient can book a doctor by specifying the patient’s location and finding the nearest chest clinics surrounding it.

Log out: Patient log out from the system.

3.1.2 DOCTORS REQUIREMENTS

Register: Register as a doctor.

Login: Login as a doctor.

Diagnose X-ray: In this function for doctors, a doctor can use plug-in offered in the portal for prediction and know if the result is positive or negative (TB).

Log out: Doctor log out from the system.

3.2 NON-FUNCTIONAL REQUIREMENTS FOR SMART (TB)

Security: To ensure that the system is secure enough from malware and attacking. Security is a measure of the system's ability to protect data and information from unauthorized access while still providing access to people and systems that are authorized. One method for thinking about system security is to think about physical security. Security tactics also leads to our four categories of tactics: detect, resist, react, and recover.

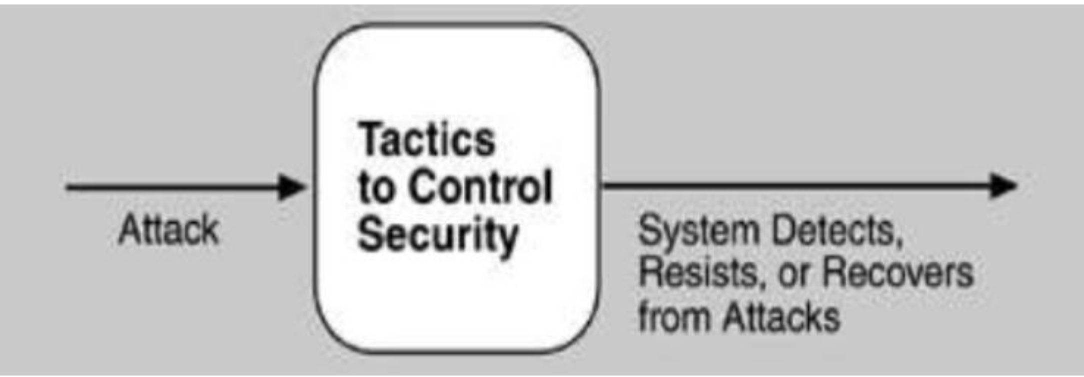

Attack
Tactics to Control Security
System Detects, Resists, or Recovers from Attacks

Figure 17: Security Tactics

Availability: To ensure that the system is working 24/7. Fundamentally, availability is about minimizing service outage time by mitigating faults. Availability tactics enable a system to endure faults so that services remain compliant with their specifications.

Fault
Tactics to Control Availability
Fault Masked or Repair Made

Figure 18: Availability Tactics

Usability: To ensure that the system is easy to use and user friendly. Also, from the Usability goals is to make an interface which any user can act on it without having an IT background.

Performance: To ensure that the system fast to navigate and respond for requests. Performance is about time and the software system's ability to meet timing requirements. Tactics to control Performance have as their goal to generate a response to an event arriving at the system within some time-based constraint.

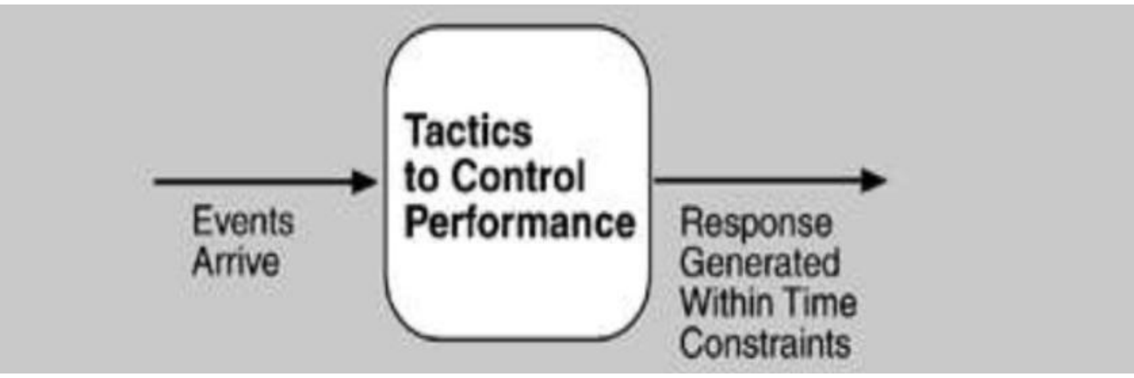


Figure 19: Performance Tactics

Interoperability: Non-functional requirement used and responsible for connecting to another systems using API. Interoperability is about the degree to which two or more systems can usefully exchange meaningful information. interoperability is not a yes-or-no proposition but has shades of meaning.

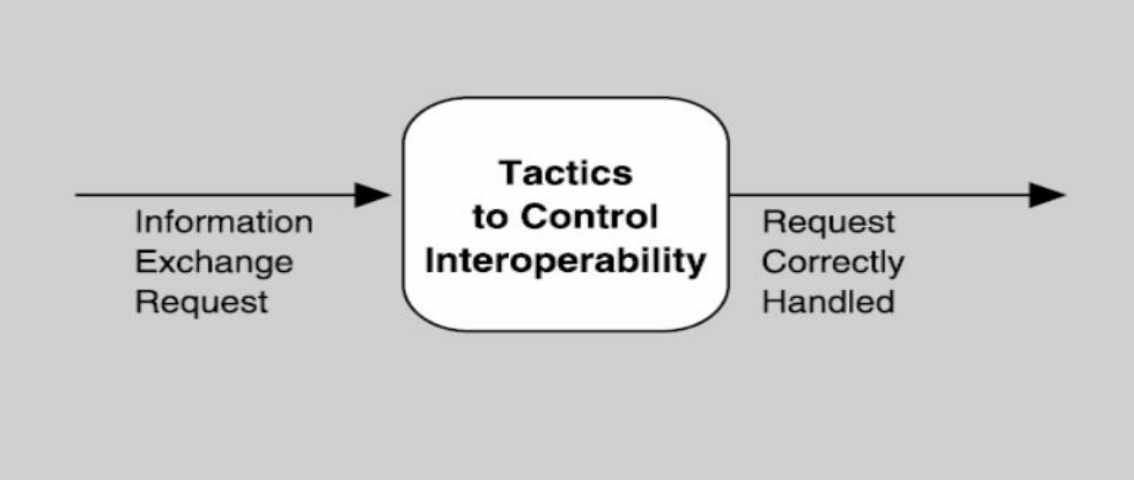


Figure 20: Interoperability Tactics

Modifiability: To ensure that the system is able to modify it, upgrade it to have new enhanced system. Modifiability Tactics to reduce the cost of making a change include making modules smaller, increasing cohesion, and reducing coupling. Deferring binding will also reduce the cost of making a change.

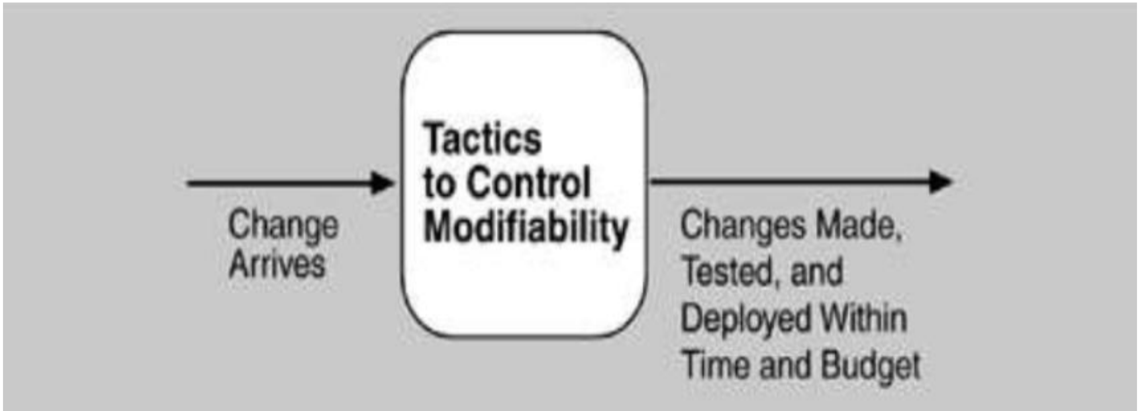
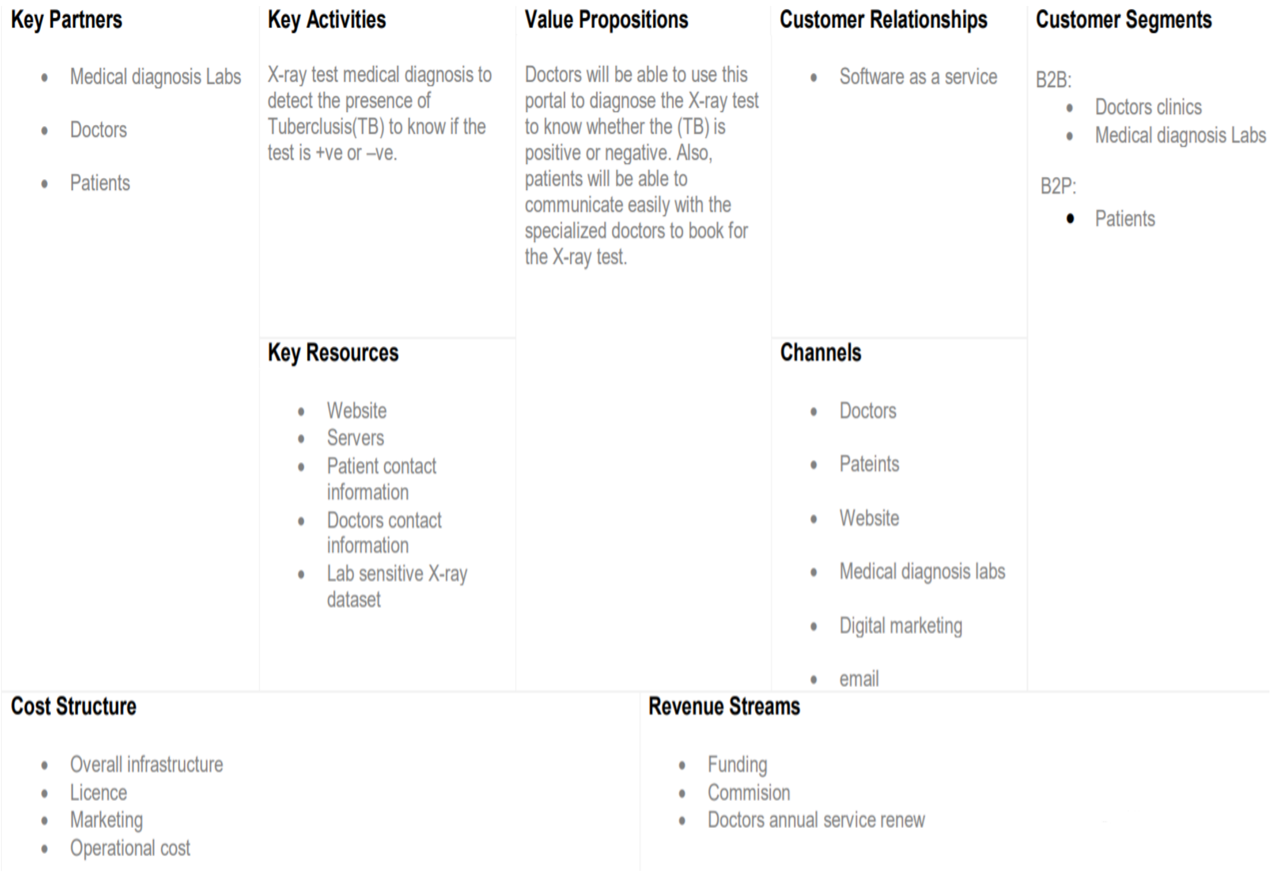

Change Arrives
Tactics to Control Modifiability
Changes Made, Tested, and Deployed Within Time and Budget

Figure 21: Modifiability Tactics

3.3 BUSINESS MODEL CANVAS



Key Partners

- Medical diagnosis Labs
- Doctors
- Patients

Key Activities

X-ray test medical diagnosis to detect the presence of Tuberclusis(TB) to know if the test is +ve or –ve.

Key Resources

- Website
- Servers
- Patient contact information
- Doctors contact information
- Lab sensitive X-ray dataset

Value Propositions

Doctors will be able to use this portal to diagnose the X-ray test to know whether the (TB) is positive or negative. Also, patients will be able to communicate easily with the specialized doctors to book for the X-ray test.

Customer Relationships

- Software as a service

Channels

- Doctors
- Pateints
- Website
- Medical diagnosis labs
- Digital marketing
- email

Customer Segments

B2B:

- Doctors clinics
- Medical diagnosis Labs

B2P:

- Patients

Cost Structure

- Overall infrastructure
- Licence
- Marketing
- Operational cost

Revenue Streams

- Funding
- Commision
- Doctors annual service renew

Figure 22: Business model canvas

3.4 DESIGN SPECIFICATION AND UMLs

3.4.1 UMLs

Use case: use case diagram is the primary form of system/software requirements for a new software program underdeveloped. Use cases specify the expected behavior (what), and not the exact method of making it happen (how). Use cases once specified can be denoted both textual and visual representation (i.e. use case diagram). A key concept of use case modeling is that it helps us design a system from the end user's perspective. It is an effective technique for communicating system behavior in the user's terms by specifying all externally visible system behavior.

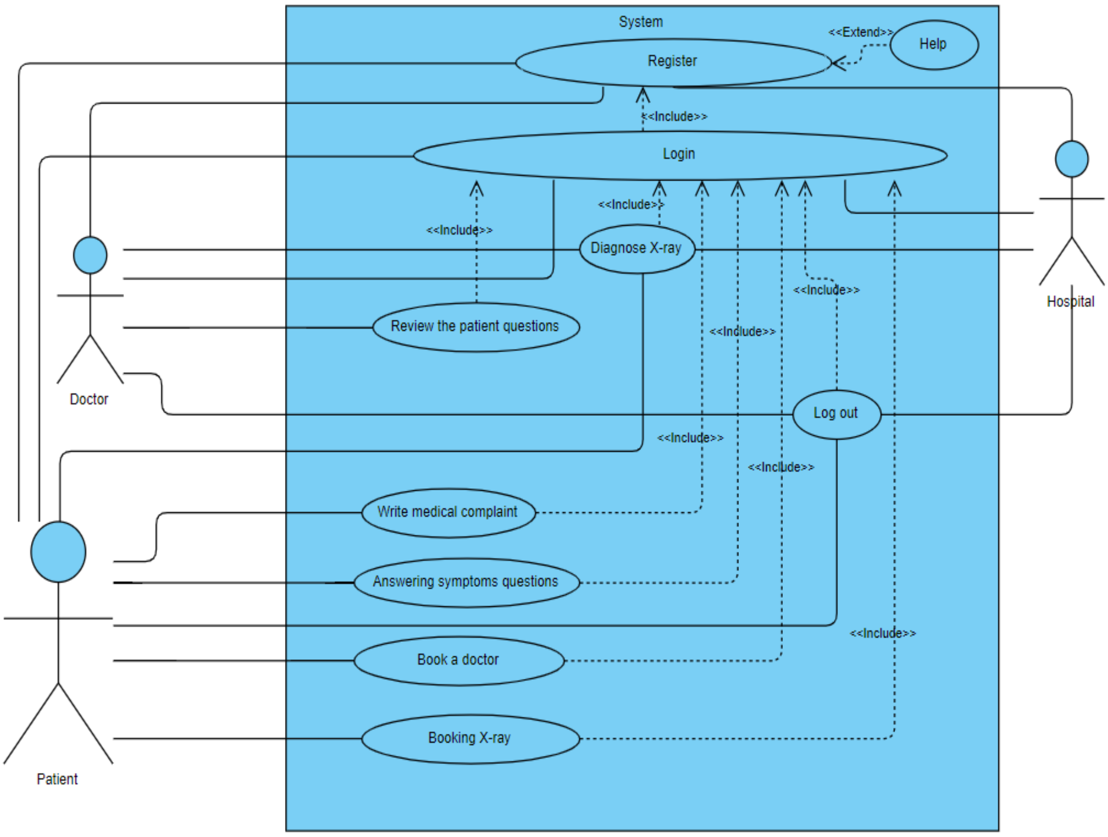


Figure 23: use case diagram

Context diagram level (0): The context diagram is used to establish the context and boundaries of the system to be modeled: which things are inside and outside of the system being modeled, and what is the relationship of the system with these external entities. A context diagram, sometimes called a level 0 data-flow diagram, is drawn in order to define and clarify the boundaries of the software system. It identifies the flows of information between the system and external entities.

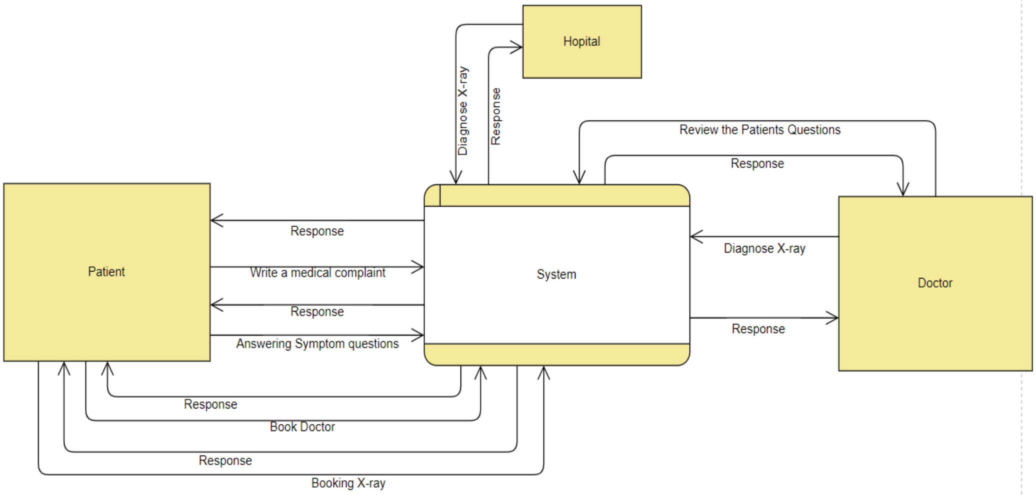


Figure 24: Context diagram level (0)

Activity diagram: Activity Diagrams describe how activities are coordinated to provide a service which can be at different levels of abstraction. Typically, an event needs to be achieved by some operations, particularly where the operation is intended to achieve a number of different things that require coordination, or how the events in a single use case relate to one another, in particular, use cases where activities may overlap and require coordination. It is also suitable for modeling how a collection of use cases coordinates to represent business workflows.

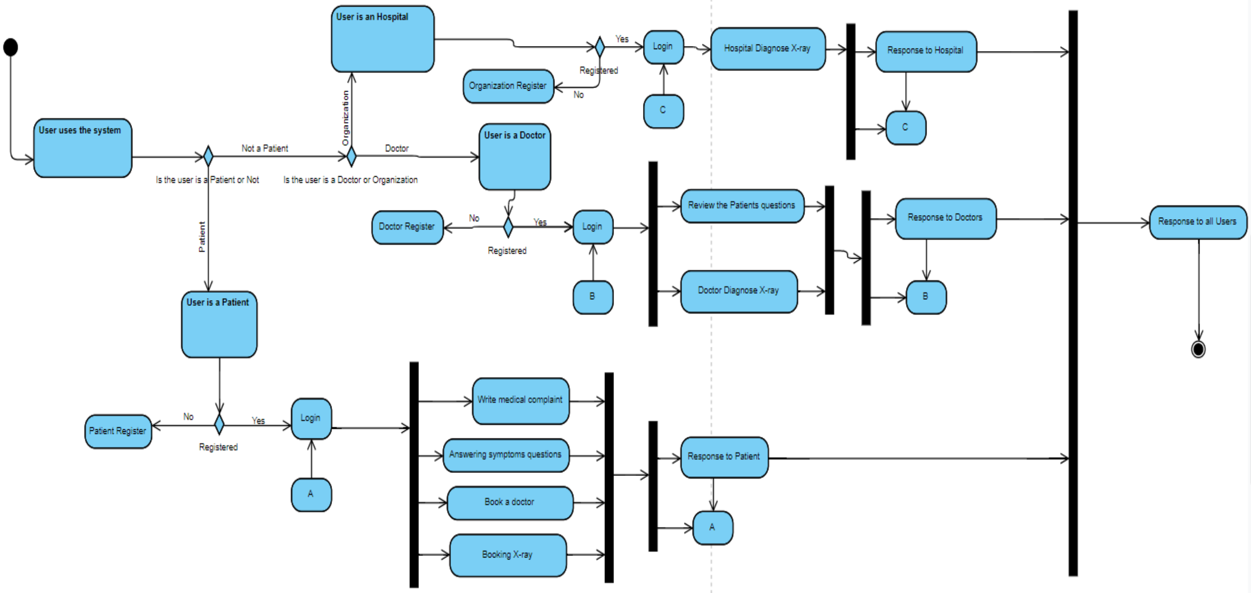


Figure 25: Activity diagram

Data flow diagram: Also known as DFD, Data flow diagrams are used to graphically represent the flow of data in a business information system. DFD describes the processes that are involved in a system to transfer data from the input to the file storage and reports generation. Data flow diagrams can be divided into logical and physical. The logical data flow diagram describes flow of data through a system to perform certain functionality of a business. The physical data flow diagram describes the implementation of the logical data flow. DFD graphically representing the functions, or processes, which capture, manipulate, store, and distribute data between a system and its environment and between components of a system.

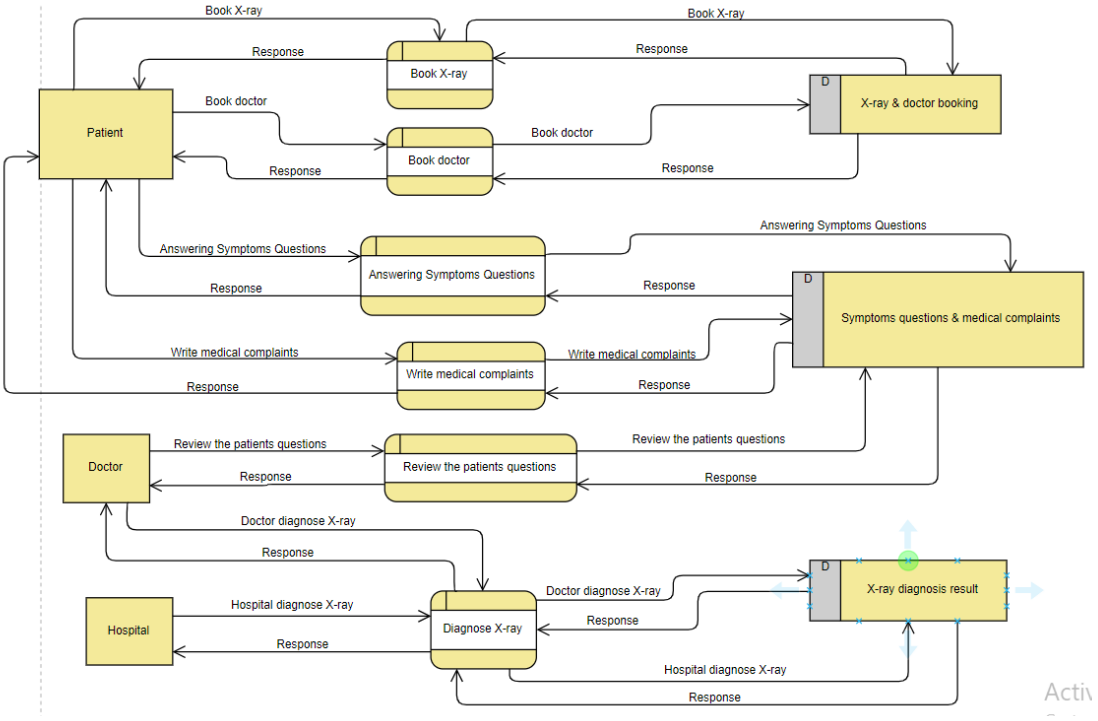


Figure 26: data flow diagram (DFD)

Class diagram: A description of a group of objects all with similar roles in the system, which consists of:

- **Structural features** (attributes) define what objects of the class "know"
    - Represent the state of an object of the class
    - Are descriptions of the structural or static features of a class
- **Behavioral features** (operations) define what objects of the class "can do"
    - Define the way in which objects may interact
    - Operations are descriptions of behavioral or dynamic features of a class

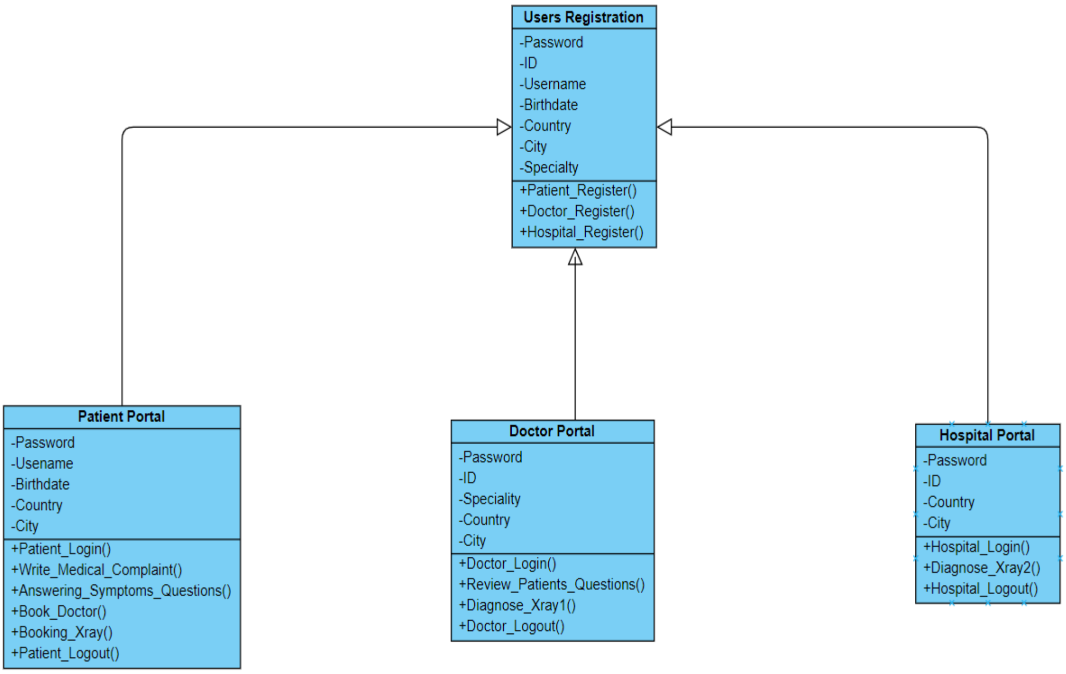


Figure 27: Class diagram

Sequence diagram: Sequence Diagrams are interaction diagrams that detail how operations are carried out. They capture the interaction between objects in the context of a collaboration. Sequence Diagrams are time focus and they show the order of the interaction visually by using the vertical axis of the diagram to represent time what messages are sent and when.

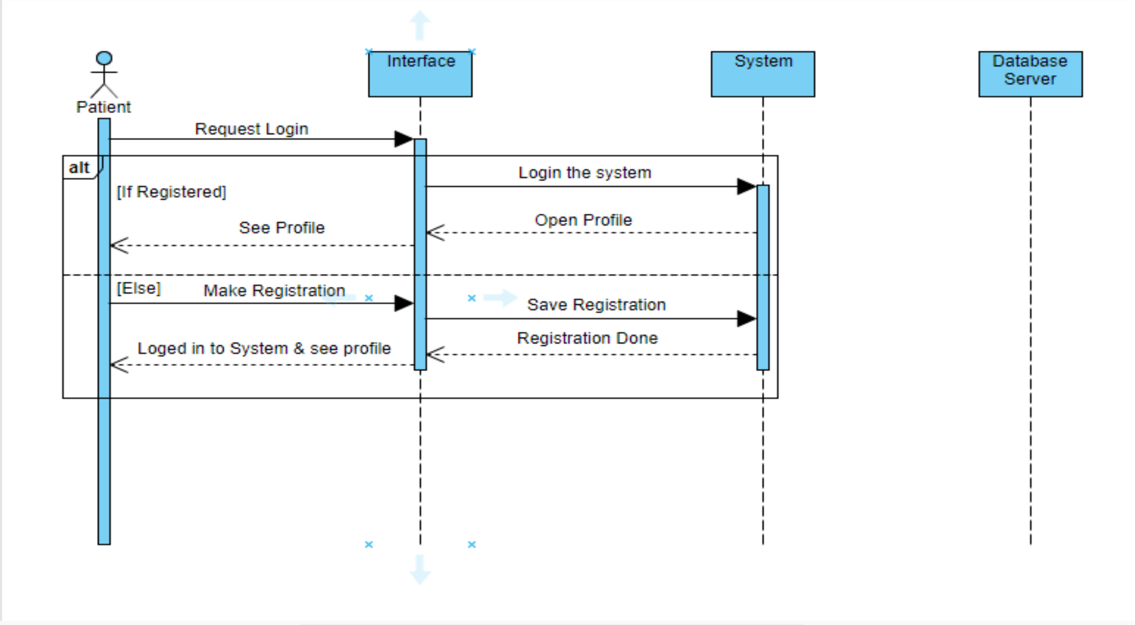


Figure 28: Patient registration sequence diagram

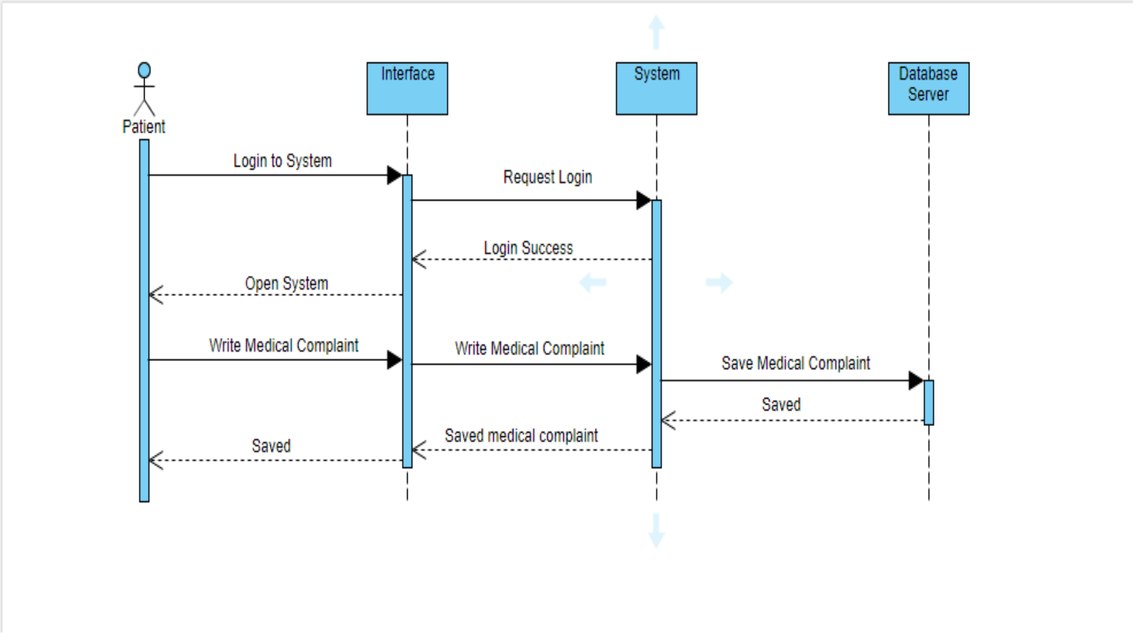


Figure 29: patients login sequence diagram

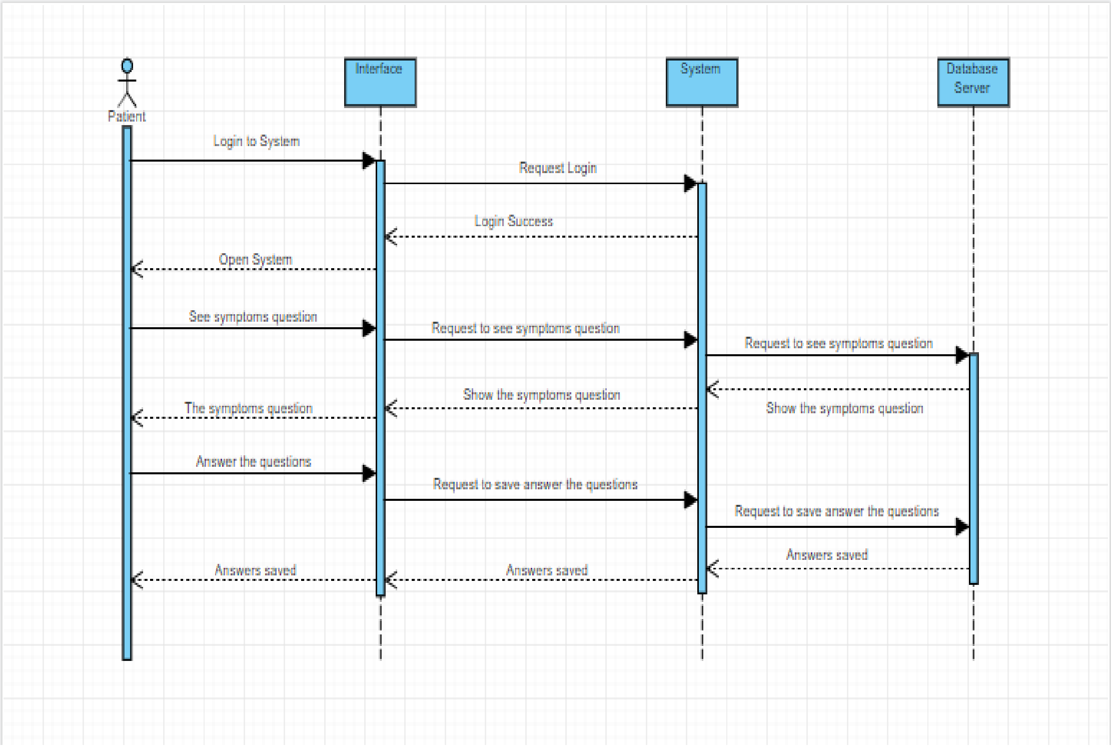


Figure 30: patients answering for symptoms questions sequence diagram

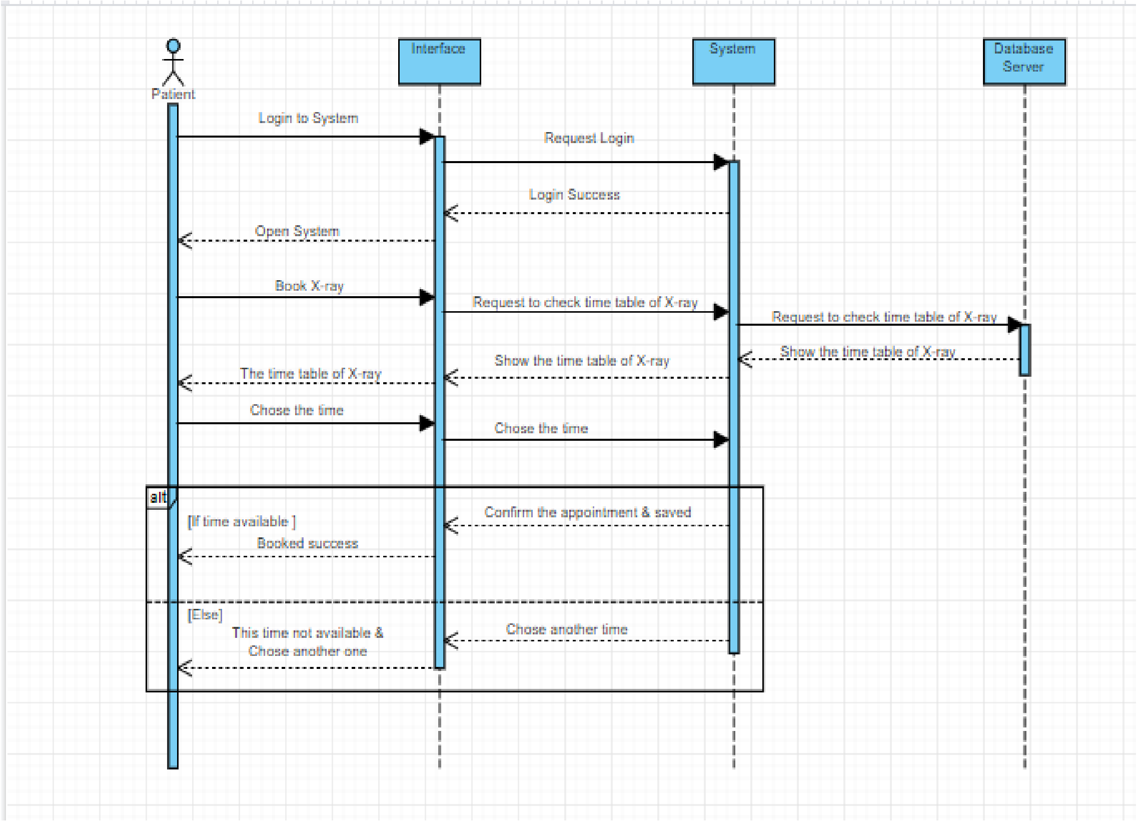


Figure 31: Book x-ray

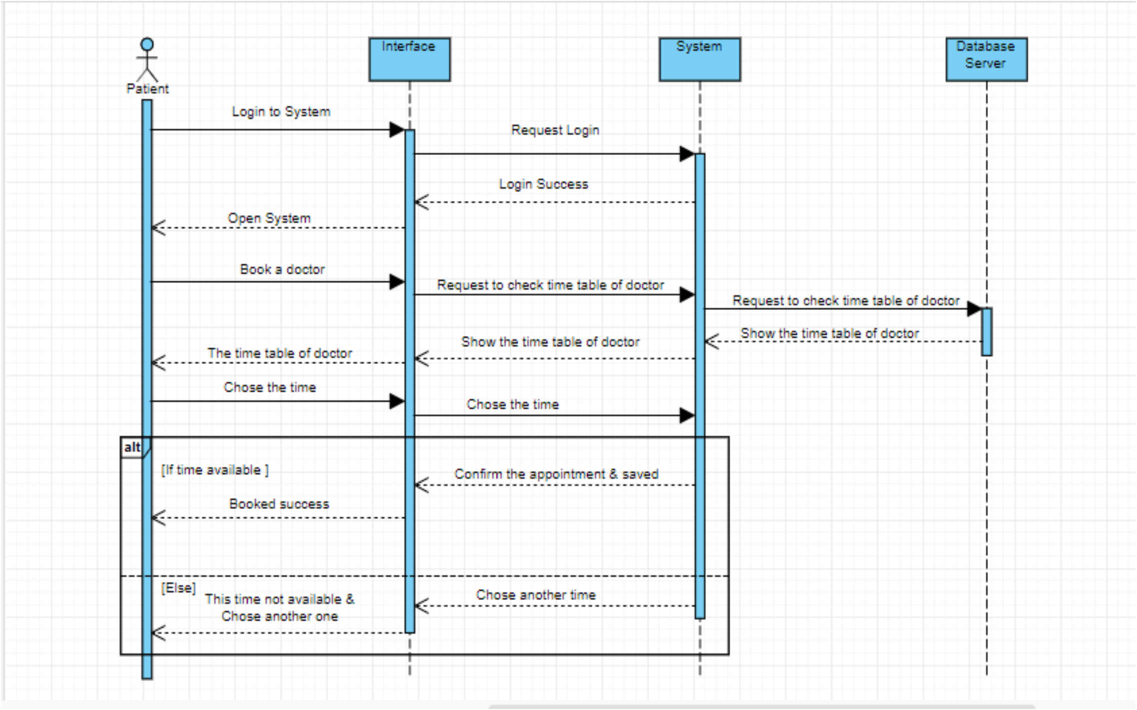


Figure 32: Book doctors

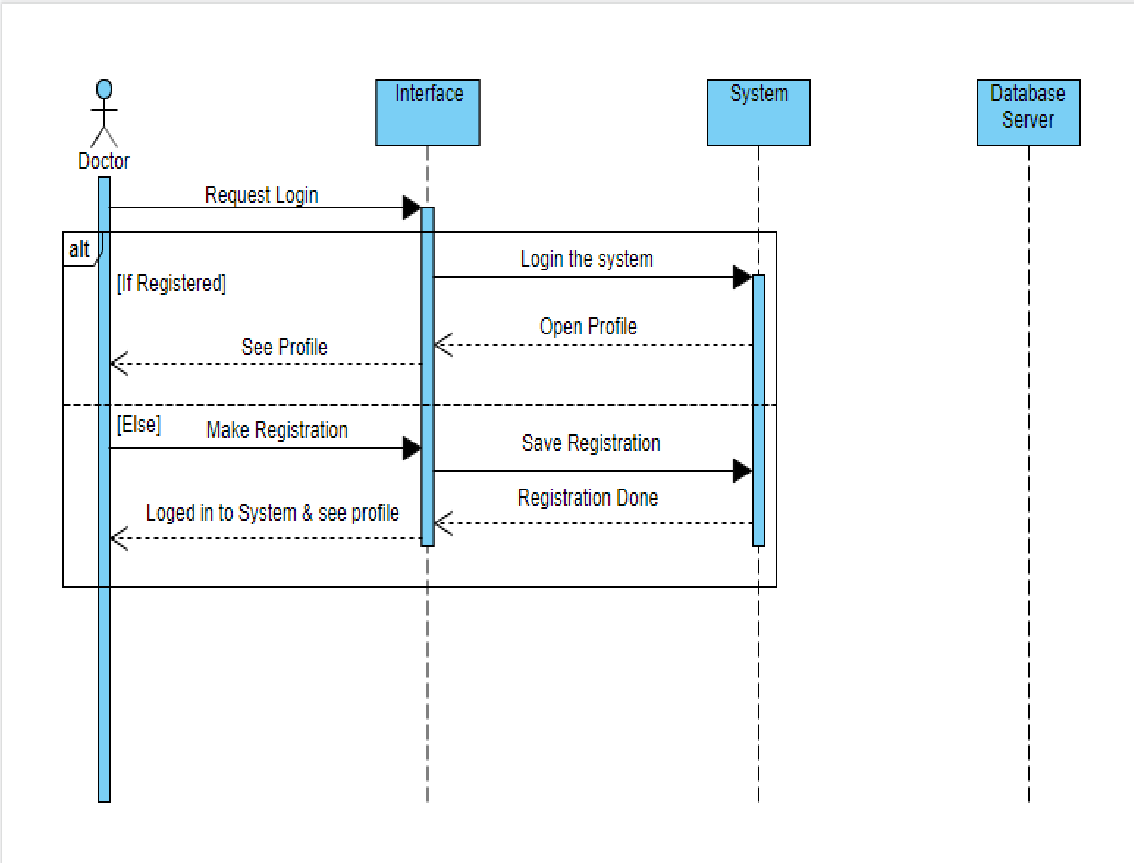


Figure 33: Doctors see patient cases

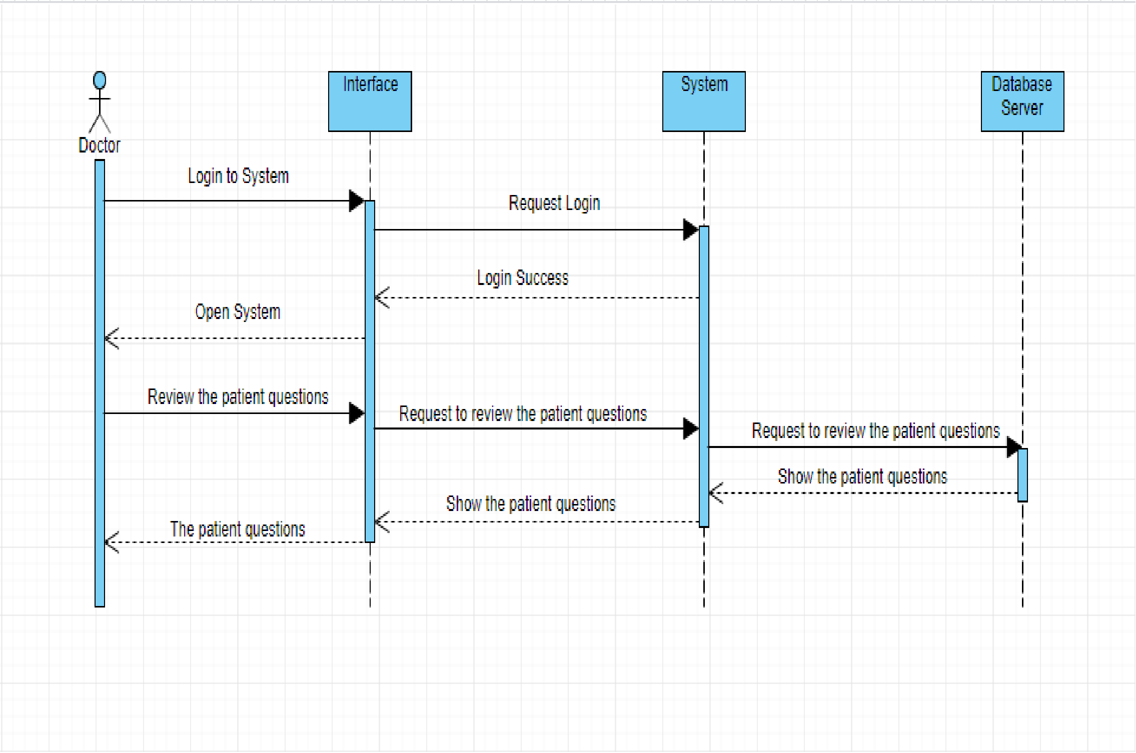

Doctor
Interface
System
Database Server
Login to System
Request Login
Login Success
Open System
Review the patient questions
Request to review the patient questions
Request to review the patient questions
Show the patient questions
Show the patient questions
The patient questions

Figure 34: Doctor review question

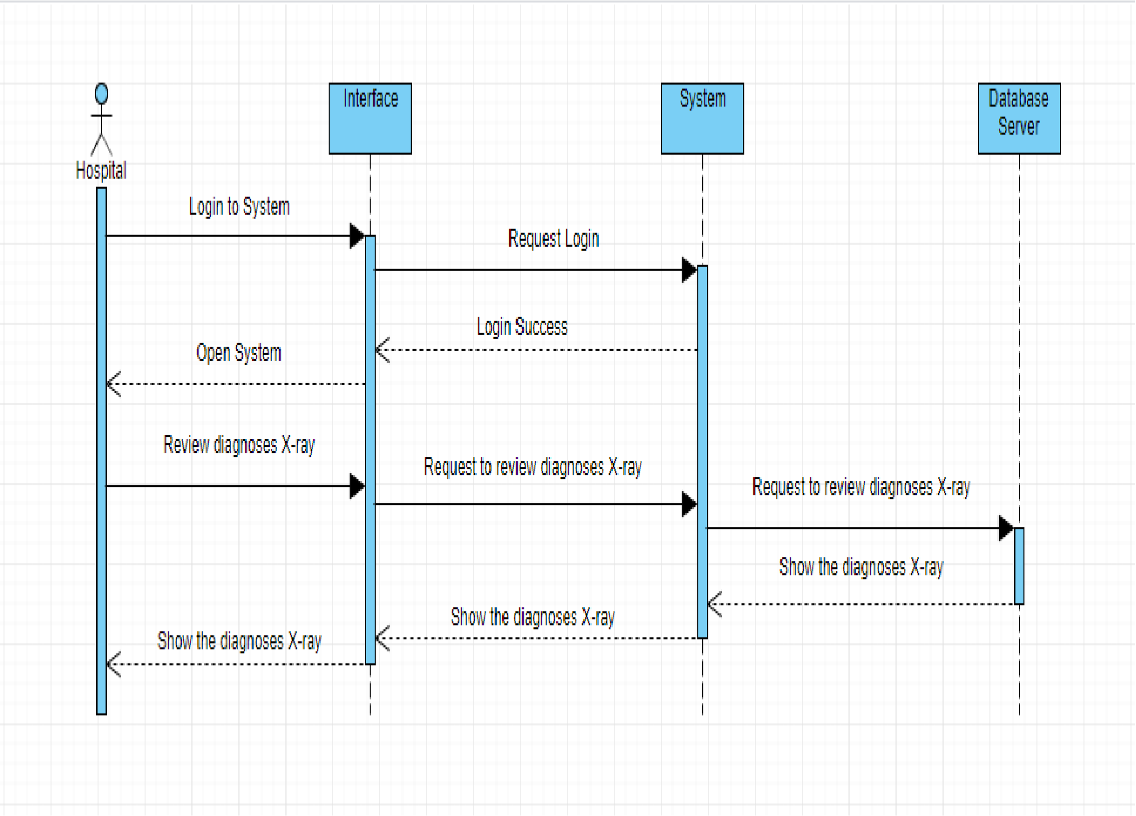


Figure 35: Doctors review result

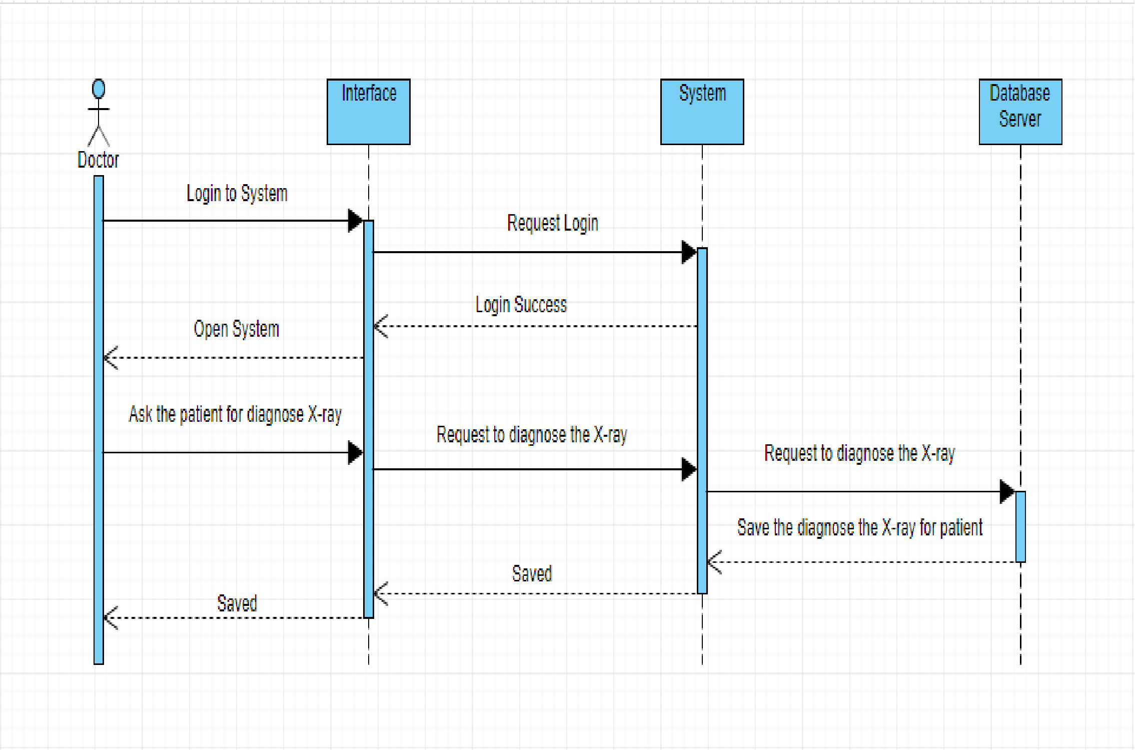

Doctor
Interface
System
Database Server
Login to System
Request Login
Login Success
Open System
Ask the patient for diagnose X-ray
Request to diagnose the X-ray
Request to diagnose the X-ray
Save the diagnose the X-ray for patient
Saved
Saved

Figure 36: Doctor take patient x-ray to diagnose it

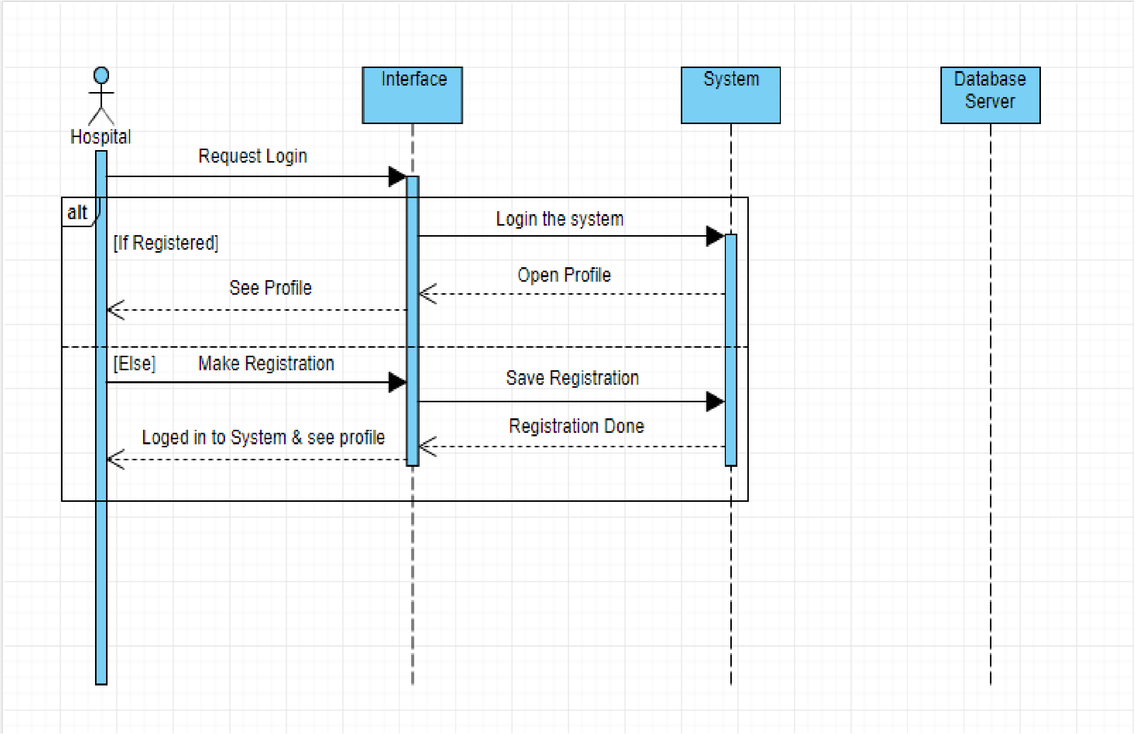


Figure 37: Hospital use Smart (TB)

3.4.2 ERD

ERD: Entity Relationship Diagram, also known as ERD, ER Diagram or ER model, is a type of structural diagram for use in database design. An ERD contains different symbols and connectors that visualize two important information: **The major entities within the system scope**, and the **inter-relationships among these entities**.

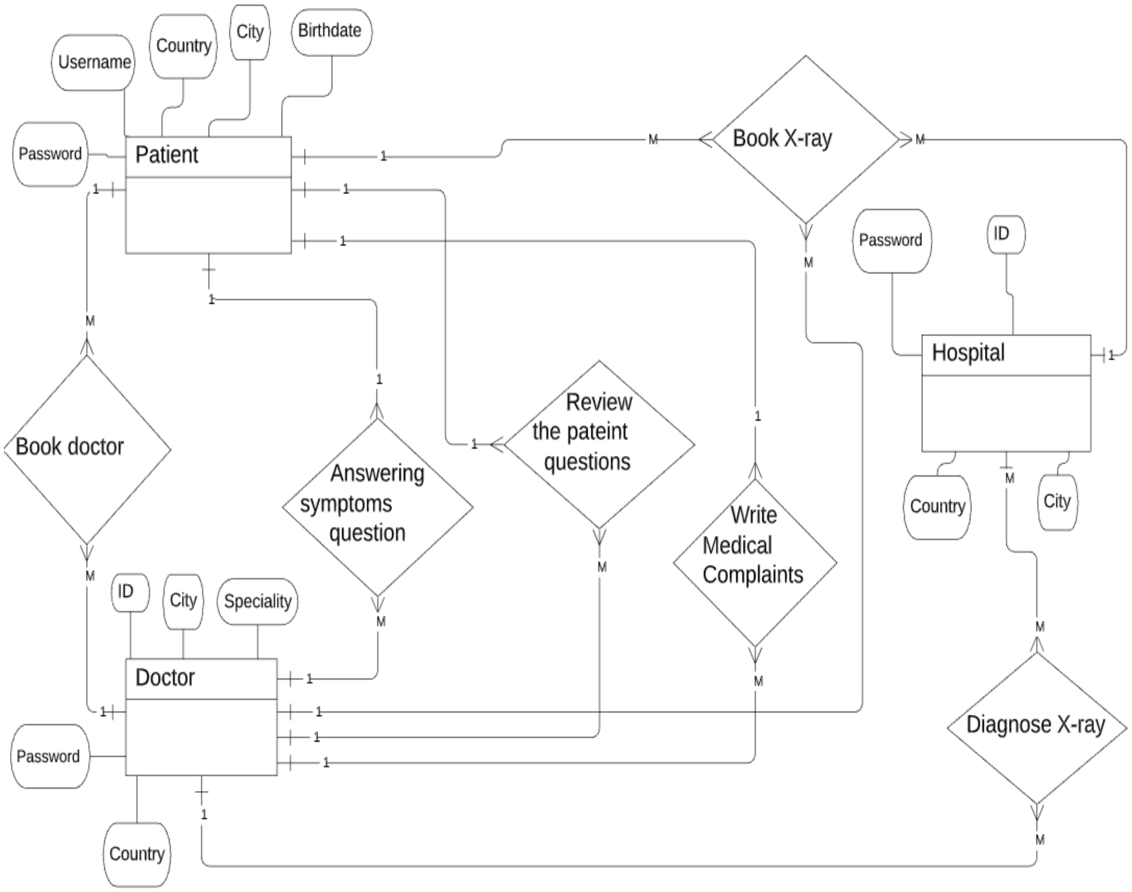
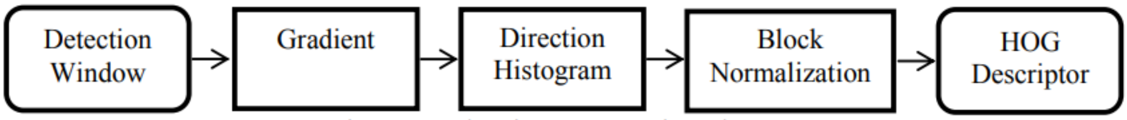


Figure 38: ERD for Smart (TB)

3.5 METHODOLOGY

3.5.1 Pre-processing: before we fed the images for CNN, we resized images to have a size of 512*512 grayscale for custom model and 224*224 (RGB) for vgg-16 or inception v3 pre-trained models. In CNNs part we don’t need to make manual feature extraction because CNN layers do the automatic feature extraction. but in our studies in tuberculosis if we pre-processed CXR images by applying histogram equalization added with segmentation for ROI will be better. our ROI in CXR images are the lungs part only. In ensemble learning there is no automatic feature extraction like deep learning (CNNs) then in our research we developed SIFT/SURF or HOG feature extraction before we fed the CXR images to the ensemble’s classifiers.

Descriptor Calculation In HOG:



3.5.2 Classification using Convolution neural network

Convolution neural networks (CNNs) are like traditional ANNs, consisting of neurons which optimize themselves by learning. Every neuron still receives an x-ray as input and performs an operation (e.g. a scalar product followed by a nonlinear function)-the basis of countless ANNs. The final layer contains class-related loss functions, and all standard tips and tricks developed for traditional ANNs remain in effect. CNN consist of many layers after the input layer, convolution layer followed by activation function, pooling layer, fully connected layer followed by activation function and the last layer is output layer and we will discuss how CNN layer working in details.

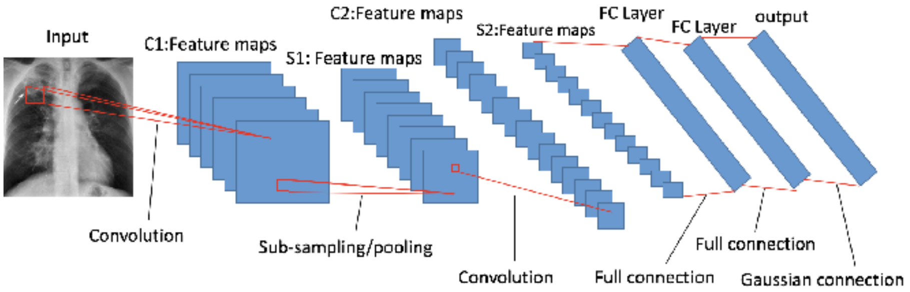


Figure 39: Example of CNN architecture

Convolution layer: the convolutional layer plays a basic role in how CNNs operate. The parameters of the layer focus around on the learnable kernels. These filters are usually small in spatial dimensionality, but spreads along the entirety of the depth of x-ray images. the layer convolves each kernel(filter) across the spatial dimension of the input to produce a 2D activation map. These activation maps can be visualised. Then, the scalar product is calculated for each value in that kernel. There are some hyperparameters in convolution layer for optimization and reduce the complexity of CNN model, the depth, the stride and zero-padding. **The depth** of the output map produced by the convolution layers and can be manually set through the number of neurons within the layer to the same region of the input. **The strides** Determines how much the window shifts in every calculation in the layer. **Zero-padding** is the basic padding method for the input border and is an efficient way of providing more dimensional control on the output map.

Kernel convolution:

$$G[m,n] = (f * h)[m,n] = \sum_{j}\sum_{k} h[j,k]f[m-j,n-k]$$

Strided Convolution:

$$n_{out} = \left\lfloor \frac{n_{in} + 2p - f}{s} + 1 \right\rfloor$$

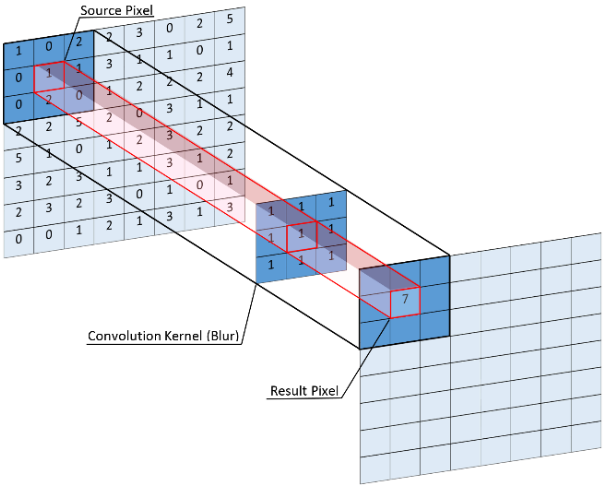


Figure 40: Kernels in Convolution lay

Activation function: After produce the new map or values produced by scaler product between kernels (filters) and input values the activation function will play a vital role after calculation produced by convolution layer the activation function will enhance some values to reduce unimportant values e.g. change negative values to be 0 and save the positive values for the next layer. The standard activation function is rectified linear unit (ReLu) and it is working (0, X). (0: for 0 and negative values) and (X: for save the new positive values to the next layer in CNN).

<u>ReLu:</u>

$$\mathrm{ReLU}(z_i) = \max(0, z_i)$$

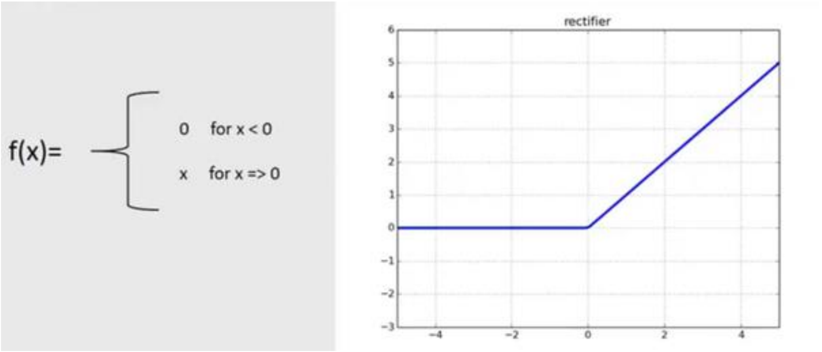


Figure 41: Rectified linear unit

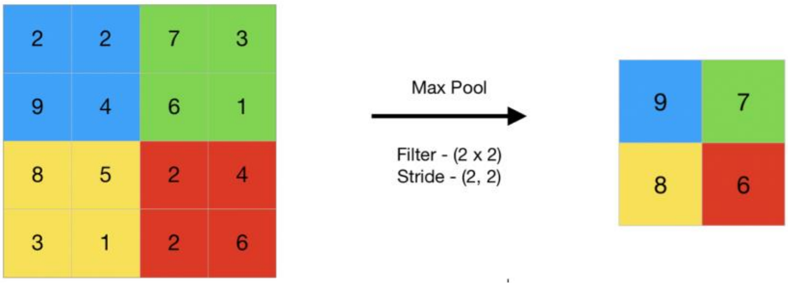
Pooling layer: Pooling layers aim to reduce the dimensionality of the convoluted input, thus further reduce the number of parameters and the computational complexity of the CNN model. The pooling layer operates over each activation map in the input, and reduce its dimensionality using the “MAX” or “AVERAGING” function. most of CNNs, they come in the form of max-pooling layers with kernels of 2×2 applied with a stride of 2 along the spatial dimensions of the input. This down the map to 25% of the original size - whilst maintaining the depth volume to its standard size.

Max-Pooling:

$$h^{l}_{xy} = \max_{i=0,..,s,j=0,..,s} \mathbf{h}^{l-1}_{(x+i)(y+j)}$$

Average-Pooling:

$$act_i = \frac{1}{M \times M} \sum_{j=1}^{M\times M} x_j$$



2	2	7	3
9	4	6	1
8	5	2	4
3	1	2	6

9	7
8	6

Figure 42: Example of Max-pooling layer

Fully-Connected layers (FC layers): FC layers composed of multiple layers and do the same task like classical shallow ANN**:**

Fully connected input layer (Flatten): takes the output from previous layer and flatten them. Flatten converts the dimension of the output to be a single vector and passes the vector as input for next layer.

The first fully-connected layer or hidden layer: takes the flattened vector from the previous FC layer and applies weights to predict the correct output(label).

Fully-connected output layer: gives the final probabilities for each label e.g. softmax, sigmoid but the most commonly used in this study is softmax which takes the highest probability from all output probabilities.

Softmax:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

Sigmoid:

$$\sigma(z_i) = \frac{1}{1+e^{-z_i}}$$

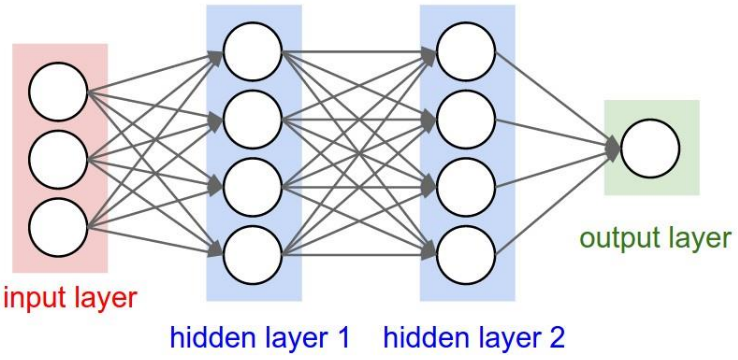


Figure 43: Example of Fully-connected layer

Fully-connected layer :

$$\begin{aligned} h_0^{out} &= I \\ h_i^{in} &= h_{i-1}^{out} * W_i + B_i \\ h_i^{out} &= \mathrm{F}_i(h_i^{in}) \end{aligned}$$

Finally, in the study most uses in tuberculosis pre-trained CNN models rather than custom models e.g. vgg-16, AlexNet most use for tuberculosis and we will discuss these models and their parameters in details in the experiments section.

3.5.3 Classification using Ensemble models

In our studies we are interested in the ensemble models. Some of ensemble models successfully get high accuracy and has the power to predict and diagnose TB correctly. We used Three main ensemble models in our research. The idea of the first method is to use pre-trained CNN as feature extractor and fed the feature vector to support vector machine for TB classification. This is idea got a good result but it is not the highest accuracy in our studies. vgg-16 feature extractor with SVM classifier achieves accuracy 90% on CXR datasets.

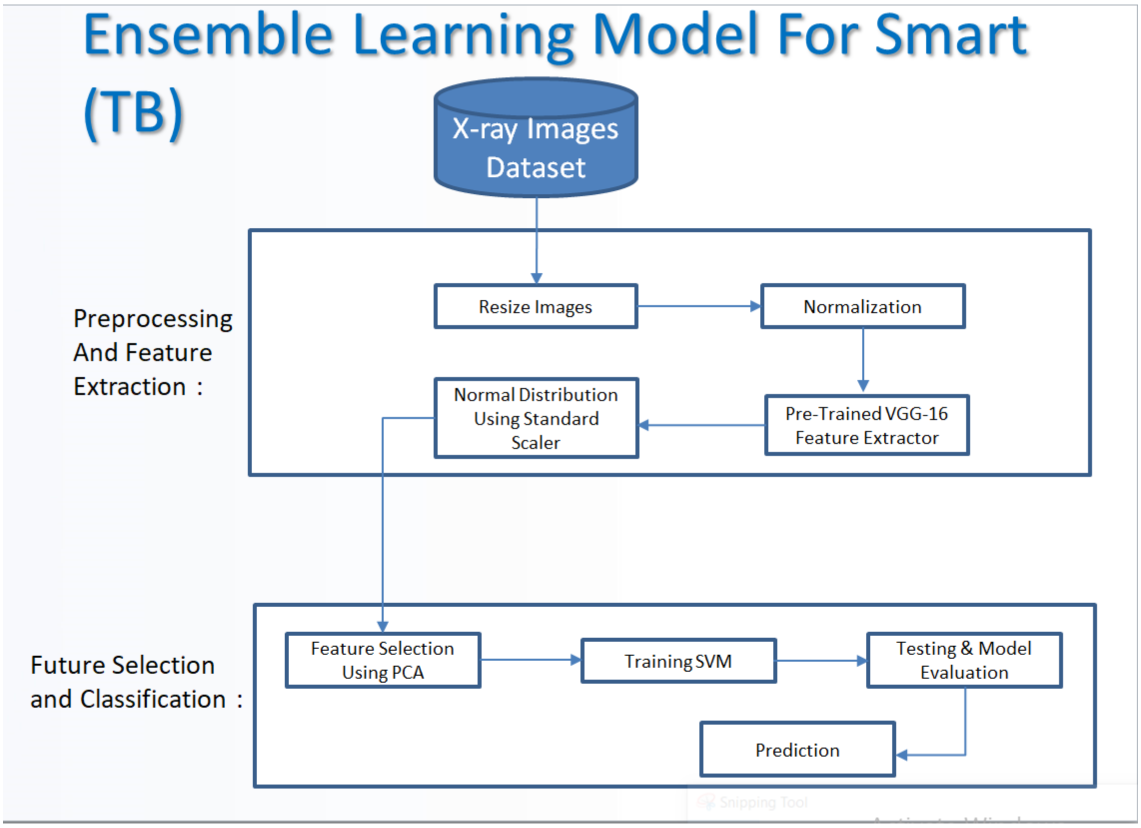

Ensemble Learning Model For Smart (TB)
X-ray Images Dataset
Preprocessing And Feature Extraction :
Resize Images
Normalization
Normal Distribution Using Standard Scaler
Pre-Trained VGG-16 Feature Extractor
Future Selection and Classification :
Feature Selection Using PCA
Training SVM
Testing & Model Evaluation
Prediction

Figure 44: Ensemble Learning Model For Smart (TB)

3.6 SYSTEM IMPLEMENTATION

3.6.1 SYSTEM IMPLEMENTATION MAP

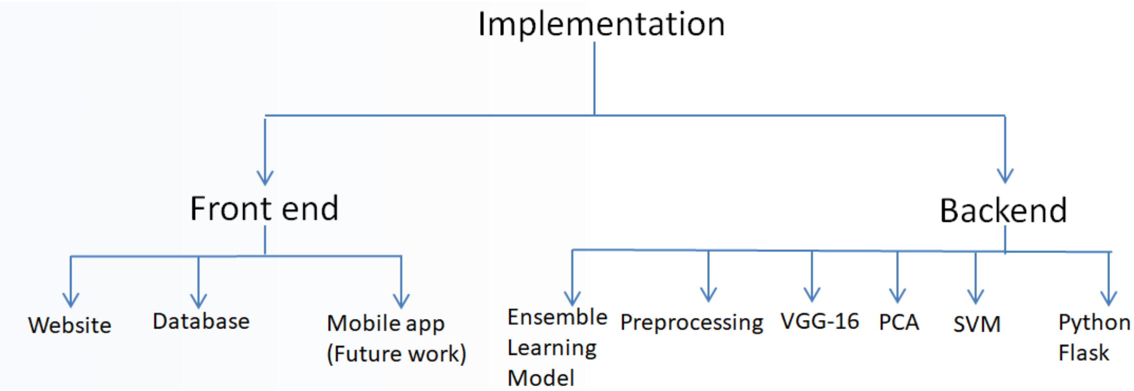
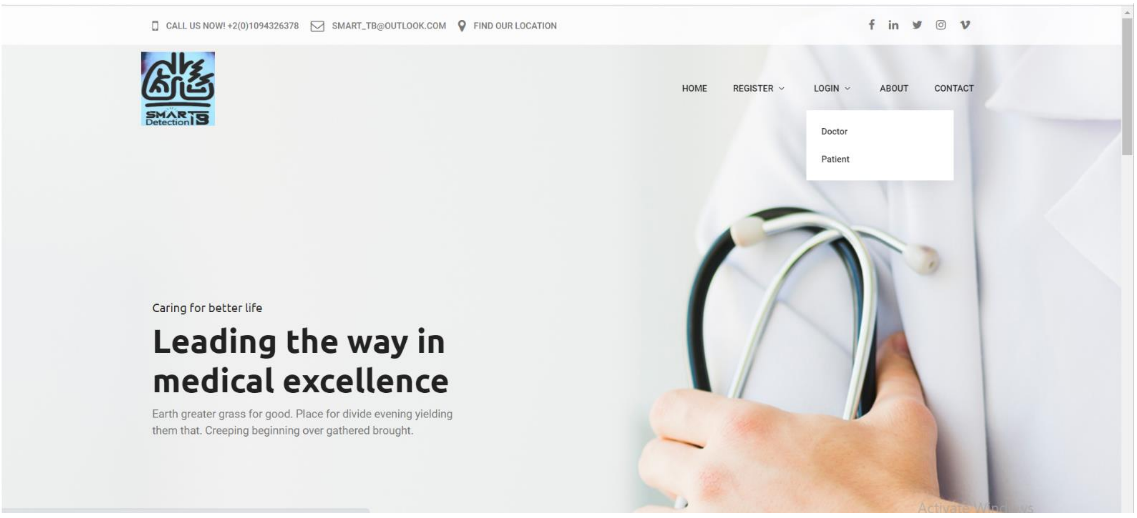


Figure 45: Implementation map

1) Front end:


CALL US NOW! +2(0)1094326378 SMART_TB@OUTLOOK.COM FIND OUR LOCATION

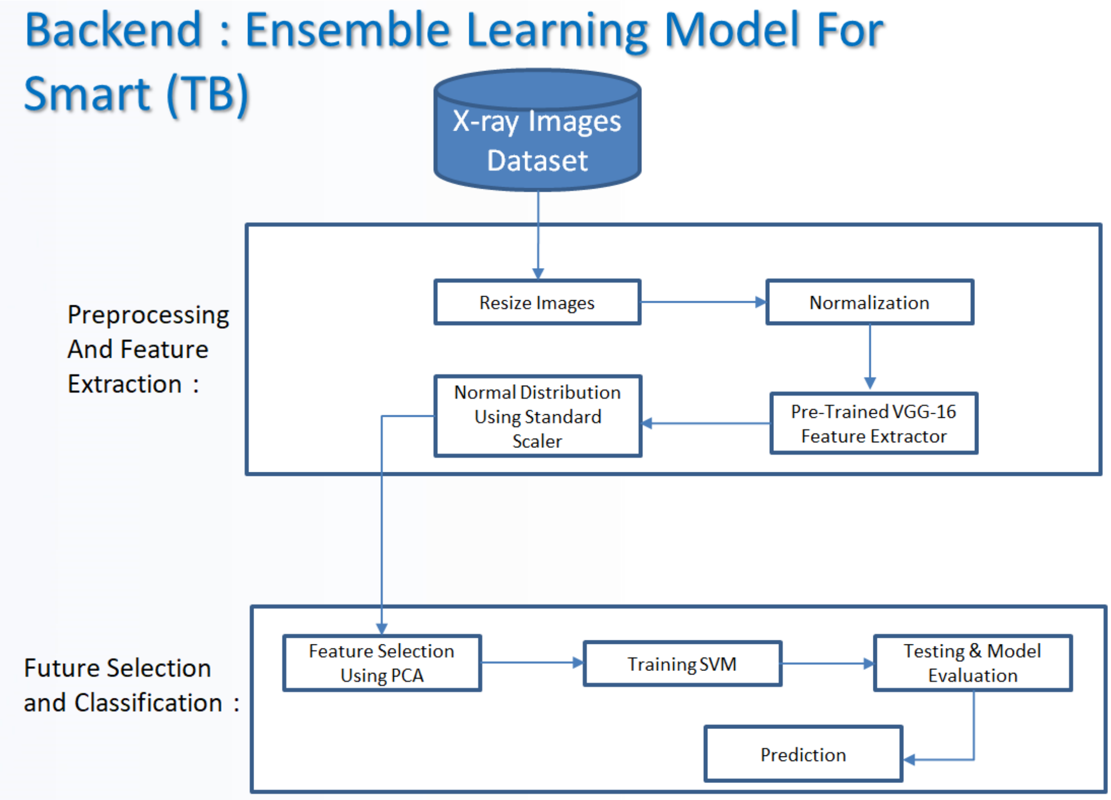
Caring for better life

Leading the way in medical excellence

Earth greater grass for good. Place for divide evening yielding them that. Creeping beginning over gathered brought.

Figure 46: Smart(TB) portal Front end

2) Backend:



Backend : Ensemble Learning Model For Smart (TB)

X-ray Images Dataset
Preprocessing And Feature Extraction :
Resize Images
Normalization
Pre-Trained VGG-16 Feature Extractor
Normal Distribution Using Standard Scaler
Future Selection and Classification :
Feature Selection Using PCA
Training SVM
Testing & Model Evaluation
Prediction

Figure 47: Smart (TB) portal Backend

3.6.2 SYSTEM IMPLEMENTATION PLATFORM

- **Backend**
    - **Hardware:**
        - **Lenovo Legon.**
        - **CPU : intel Core i7 8th generation.**
        - **GPU : NVidia GTX 1060 TI.**
    - **Operating system: Windows 10.**
    - **Tools : PyCharm , Anaconda Spyder.**
- **Frontend**
    - **Tools : PyCharm, Visual Studio Code.**

3.7 RISK ASSESMENT PLAN

THE FOLLOWING STEPS SHOWS HOW OUR RISK ASSESMENT PALN WORKS:

1 Smart(TB) detection risk assessment introduction:

Our risk assessment is about the overall risk identification process, risk analysis and risk evaluation for our project.

Risk assessment techniques used:

checklist of known threats and hazards

Table A: Risk Classifications

Risk Level	**Risk Description & Necessary Actions**
High	**The loss of confidentiality, integrity, or availability could be expected to have a severe or catastrophic adverse effect on organizational operations, organizational assets or individuals.**
Moderate	**The loss of confidentiality, integrity, or availability could be expected to have a serious adverse effect on organizational operations, organizational assets or individuals.**
Low	**The loss of confidentiality, integrity, or availability could be expected to have a limited adverse effect on organizational operations, organizational assets or individuals.**

2 Identify and Prioritize Smart(TB) portal Assets

Table B: Smart(TB) portal assets

Priority	Asset	Asset information
2)	Website (Critical)	• Software • Technical security controls • Interfaces
1)	Servers (Critical)	• Information storage protection • IT Security architecture • Technical security controls
4)	Patient contact information (Low)	• Data • Users
5)	Doctors contact information (Medium)	• Data • Users
3)	Lab sensitive X-ray dataset (Critical)	• Data

3 Identify Threats

The threats identified are listed in Table C.

Table C: Threats Identified

Threat	Threat type(Category)	Threat identity
IP address spoofing to the servers	Malicious humans	Interception
Remote code execution to the website	Malicious humans	Interception
Data manipulation to the Hospital lab sensitive X-ray dataset	Accidental human interference	make mistakes e.g. (Accidentally deleting important files)
Sql Injection for Patient contact information	Malicious humans	Interception
Sql Injection for Doctors contact information	Malicious humans	Interception

4 Identify Vulnerabilities

The way vulnerabilities combine with credible threats to create risks is identified Table D.

Table D: Vulnerabilities, Threats, and Risks

Risk No.	Vulnerability	Threat	Risk of Compromise of	Risk Summary
1	***No Use of cryptographic network protocols***	*IP address spoofing to the servers*	**user data or spreading malware to harm users accounts**	*The attacker motivation is to launch attacks against network hosts, steal data, spread malware or bypass access controls*
2	***NO timely installation of software update***	Remote code execution *to the website*	***Website performance and confidentiality***	*When the attacker gain a user administrative access, the attacker can do any fraud or any other illegal actions.*
3	***NO regularly back up to our data***	*Data manipulation to the Hospital lab sensitive X-ray dataset*	***Hospital lab sensitive X-ray dataset***	*Attacker will fraud the datasets so may an X-ray test result be wrong because of the inaccurate prediction due to the false dataset.*
4	***NO Parameterized queries in the portal system***	Sql Injection *for Patient contact information*	***Patient account personal data***	*Attacker motivation is to know secret info. of Patient to override their valuable data, or even to execute dangerous system level commands on the database host.*

Risk No.	Vulnerability	Threat	Risk of Compromise of	Risk Summary
5	***NO Parameterized queries in the portal system***	Sql Injection *for Doctors contact information*	***Doctor account personal data***	*Attacker motivation is to know secret info. of Doctor to override their valuable data, or even to execute dangerous system level commands on the database host.*

5 Analyze Controls

Table E documents the IT security controls in place and planned for the IT system.

Table E: Security Controls

Control Area	In-Place/ Planned	Description of Controls
1 Risk Management		
1.1 IT Security Roles & Responsibilities	Planned	***Owning SSL certification for our web application & using cloud services afford it to us.***
1.2 Business Impact Analysis	Planned	***Our business core idea and goals was discussed before, but in case of the lack of availability backups and fail over strategies will help us mitigates our technical problems to continue our business.***
1.3 IT System & Data Sensitivity Classification	Planned	***Owning SSL certification for our web application & using cloud services afford it to us.***
1.4 IT System Inventory & Definition	Planned	***Owning SSL certification for our web application & using cloud services afford it to us.***
1.5 Risk Assessment	In-Place	***After identifying our vulnerabilities we are now planning to control our risk assessment.***
1.6 IT Security Audits	Planned	***Owning SSL certification for our web application & using cloud services afford it to us.***
2 IT Contingency Planning		
2.1 Continuity of Operations Planning	Planned	***Owning SSL certification for our web application & using cloud services afford it to us.***

Control Area	In-Place/ Planned	Description of Controls
2.2 IT Disaster Recovery Planning	Planned	***Owning SSL certification for our web application & using cloud services afford it to us.***
2.3 IT System & Data Backup & Restoration	Planned	***Owning SSL certification for our web application & using cloud services afford it to us.***
	3 IT Systems Security	
3.1 IT System Hardening	Planned	***Smart(TB) portal doesn't contain hardware as a part of it as it is a medical diagnosis system depending on cloud services.***
3.2 IT Systems Interoperability Security	In-Place	***After identifying our vulnerabilities we are now planning to control our risk assessment to protect*** **Interoperability quality attribute.**
3.3 Malicious Code Protection	In-Place	***We are planning to timely install software update, this ranks as the top cybersecurity measure in preventing remote or malicious code execution attacks.***
3.4 IT Systems Development Life Cycle Security	***In-Place***	***We are planning to apply a suitable framework e.g. (ITIL framework).***
	4 Logical Access Control	
4.1 Account Management	***Planned***	***It was planned in our prototype***
4.2 Password Management	***Planned***	***It was planned in our prototype***
4.3 Remote Access	***Planned***	***It was planned in our prototype***
	5 Data Protection	
4.4 Data Storage Media Protection	***Planned***	***Owning SSL certification for our web application & using cloud services afford it to us.***
4.5 Encryption	***Planned***	***Owning SSL certification for our web application & using cloud services afford it to us.***
	6 Facilities Security	
6.1 Facilities Security	***Planned***	***Owning SSL certification for our web application & using cloud services afford it to us.***
	7 Personnel Security	

Control Area	In-Place/ Planned	Description of Controls
7.1 Access Determination & Control	***Planned***	***Owning SSL certification for our web application & using cloud services afford it to us.***
7.2 IT Security Awareness & Training	***In-Place***	***We are planning to design tutorials to all our portal users to know well how to use it and protect well their info. .***
7.3 Acceptable Use	***Planned***	***Owning SSL certification for our web application & using cloud services afford it to us.***
8 Threat Management		
8.1 Threat Detection	***Planned***	***Owning SSL certification for our web application & using cloud services afford it to us.***
8.2 Incident Handling	***In-Place***	***We are handling it in our security risk assessment.***
8.3 Security Monitoring & Logging	***Planned***	***Owning SSL certification for our web application & using cloud services afford it to us.***
9 IT Asset Management		
9.1 IT Asset Control	***In-Place***	***We are planning to many controls to handle our IT Asset***
9.2 Software License Management	***Planned***	**Software License for users is for free until now and for us we manage our services availability by renewing our ownership to the *SSL certification for our web application & our cloud services.***
9.3 Configuration Management & Change Control	***Planned***	***This configuration criteria changes according to the change of our business criteria.***

Table E correlates the risks identified in Table C with relevant IT security controls documented in Table D and with other mitigating or exacerbating factors.

Table F: Risks-Controls-Factors Correlation

Risk No.	**Risk Summary**	**Correlation of Relevant Controls & Other Factors**
1	*The attacker motivation is to launch attacks against network hosts, steal data, spread malware or bypass access controls*	Controls should be preventive
2	*When the attacker gain a user administrative access, the attacker can do any fraud or any other illegal actions.*	Controls should be preventive
3	*Attacker will fraud the datasets so may an X-ray test result be wrong because of the inaccurate prediction due to the false dataset.*	Controls should be preventive
4	*Attacker motivation is to know secret info. of Patient to override their valuable data, or even to execute dangerous system level commands on the database host.*	Controls should be detective
5	*Attacker motivation is to know secret info. of Doctor to override their valuable data, or even to execute dangerous system level commands on the database host.*	Controls should be detective

6 Determine the Likelihood of an Incident

Table G defines the risk likelihood ratings.

Table G: Risk Likelihood Definitions

Effectiveness of Controls	Probability of Threat Occurrence (Natural or Environmental Threats) or Threat Motivation and Capability (Human Threats)		
	Low	Moderate	High
Low	Medium	High	High
Moderate	Low	Medium	High
High	Low	Low	Medium

Table G, evaluates the effectiveness of controls and the probability or motivation and capability of each threat to BFS and assigns a likelihood, as defined in Table F, to each risk documented in Table C.

Table H: Risk Likelihood Ratings

Risk No.	Risk Summary	Risk Likelihood Evaluation	Risk Likelihood Rating
1	*The attacker motivation is to launch attacks against network hosts, steal data, spread malware or bypass access controls*	**High**	**High**
2	*When the attacker gain a user administrative access, the attacker can do any fraud or any other illegal actions.*	**High**	**High**
3	*Attacker will fraud the datasets so may an X-ray test result be wrong because of the inaccurate prediction due to the false dataset.*	**High**	**High**
4	*Attacker motivation is to know secret info. of Patient to override their valuable data, or even to execute dangerous system level commands on the database host.*	**Low**	**Low**
5	*Attacker motivation is to know secret info. of Doctor to override their valuable data, or even to execute dangerous system level commands on the database host.*	**Medium**	**Medium**

7 Assess the Impact a Threat Could Have

Table I documents the ratings used to evaluate the impact of risks.

Table I: Risk Impact Rating Definitions

Magnitude of Impact	Impact Definition
High	**Occurrence of the risk: (1) May result in human death or serious injury; (2) May result in the loss of major tangible assets, resources or sensitive data; or (3) May significantly harm, or impede the mission, reputation or interest.**
Medium	**Occurrence of the risk: (1) May result in human injury; (2) May result in the costly loss of tangible assets or resources; or (3) May violate, harm, or impede the mission, reputation or interest.**
Low	**Occurrence of the risk: (1) May result in the loss of some tangible assets or resources or (2) May noticeably affect the mission, reputation or interest.**

Table J documents the results of the impact analysis, including the estimated impact for each risk identified in Table D and the impact rating assigned to the risk.

Table J: Risk Impact Analysis

Risk No.	Risk Summary	Risk Impact	Risk Impact Rating
1	*The attacker motivation is to launch attacks against network hosts, steal data, spread malware or bypass access controls*	**High**	**High** ***(2)***
2	*When the attacker gain a user administrative access, the attacker can do any fraud or any other illegal actions.*	**High**	**High** ***(3)***

Risk No.	Risk Summary	Risk Impact	Risk Impact Rating
3	*Attacker will fraud the datasets so may an X-ray test result be wrong because of the inaccurate prediction due to the false dataset.*	**High**	**High** ***(1)***
4	*Attacker motivation is to know secret info. of Patient to override their valuable data, or even to execute dangerous system level commands on the database host.*	**Low**	**Low** ***(2)***
5	*Attacker motivation is to know secret info. of Doctor to override their valuable data, or even to execute dangerous system level commands on the database host.*	**Medium**	**Medium** ***(1)***

8 Prioritize the Information Security Risks

Table K documents the criteria used in determining overall risk ratings.

Table K: Overall Risk Rating Matrix

Risk Likelihood	Risk Impact		
	Low (10)	**Moderate (50)**	**High (100)**
High (1.0)	**Low 10 x 1.0 = 10**	**Moderate 50 x 1.0 = 50**	**High 100 x 1.0 = 100**
Moderate (0.5)	**Low 10 x 0.5 = 5**	**Moderate 50 x 0.5 = 25**	**Moderate 100 x 0.5 = 50**
Low (0.1)	**Low 10 x 0.1 = 1**	**Low 50 x 0.1 = 5**	**Low 100 x 0.1 = 10**

Risk Scale: Low (1 to 10); Moderate (>10 to 50); High (>50 to 100)

Table L assigns an overall risk rating, as defined in Table K, to each of the risks documented in Table D.

Table L: Overall Risk Ratings Table

Risk No.	Risk Summary	Risk Likelihood Rating	Risk Impact Rating	Overall Risk Rating
1	*The attacker motivation is to launch attacks against network hosts, steal data, spread malware or bypass access controls*	**High**	**High**	**100 x 1.0 = 100**
2	*When the attacker gain a user administrative access, the attacker can do any fraud or any other illegal actions.*	**High**	**High**	**100 x 1.0 = 100**
3	*Attacker will fraud the datasets so may an X-ray test result be wrong because of the inaccurate prediction due to the false dataset.*	**High**	**High**	**100 x 1.0 = 100**
4	*Attacker motivation is to know secret info. of Patient to override their valuable data, or even to execute dangerous system level commands on the database host.*	**Low**	**Low**	**10 x 0.1 = 1**
5	*Attacker motivation is to know secret info. of Doctor to override their valuable data, or even to execute dangerous system level commands on the database host.*	**Medium**	**Medium**	**50 x 0.5 = 25**

9 Recommend Controls

Table M documents recommendations for the risks identified in Table D.

Table M: Recommendations

Risk No.	Risk	Risk Rating	Recommendations
1	**May result in the loss of major assets, resources or sensitive data**	**100**	**A mitigation activity for this threat is Use cryptographic network protocols: Transport Layer Security (TLS), Secure Shell (SSH), HTTP Secure (HTTPS) and other secure communications protocols bolster spoofing attack prevention efforts by encrypting data before it is sent and authenticating data as it is received.**
2	**May significantly harm, or impede the Portal mission, reputation or interest.**	**100**	***Timely patching or timely installation of software update ranks as the top cybersecurity measure in preventing remote code execution attacks. This the best mitigation activity for that threat.***
3	**May result in human death or serious health problems**	**100**	***we will use Jupyter Notebooks, the popular environment among data scientists, to predict the salaried class using both raw data and privacy-protected data. We will be using CryptoNumerics' privacy libraries for the privacy algorithms and sklearn for the regression.***
4	**May noticeably affect the Portal mission, reputation or interest.**	**1**	***A mitigation activity for this threat are Parameterized queries which are simple to write and understand. They force you to define the SQL query and use placeholders for user-provided variables in the query. After the SQL statement is defined, you can pass each parameter to the query. This allows the database to distinguish between the SQL command and data supplied by a user. If you properly parametrize SQL queries, all user input that is passed to the database is treated as data and can never be confused as being part of a command.***

5	**May result in serious health problems**	**25**	***A mitigation activity for this threat are Parameterized queries which are simple to write and understand. They force you to define the SQL query and use placeholders for user-provided variables in the query. After the SQL statement is defined, you can pass each parameter to the query. This allows the database to distinguish between the SQL command and data supplied by a user. If an attacker inputs SQL commands, the parameterized query treats them as untrusted input and the database does not execute injected SQL commands.***

10 Document the Results

Exhibit 1: Risk Assessment Matrix

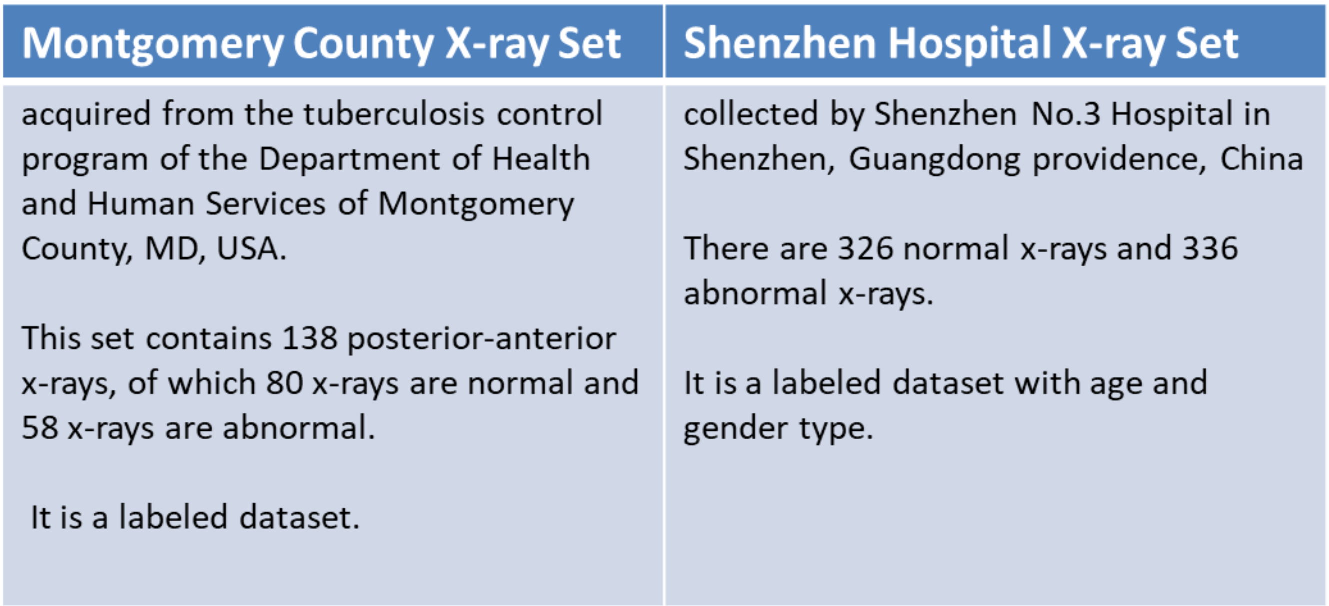
Risk No.	Vulnerability	Threat	Risk	Risk Summary	Risk Likelihood Rating	Risk Impact Rating	Overall Risk Rating	Analysis of Relevant Controls and Other Factors
1	***No Use of cryptographic network protocols***	**IP address spoofing to the servers**	**May result in the loss of major assets, resources or sensitive data**	*The attacker motivation is to launch attacks against network hosts, steal data, spread malware or bypass access controls*	**High**	**High**	**100**	Controls should be preventive

2	***NO timely installation of software update***	**Remote code execution to the website**	**May significantly harm, or impede the Portal mission, reputation or interest.**	*When the attacker gain a user administrative access, the attacker can do any fraud or any other illegal actions.*	**High**	**High**	**100**	Controls should be preventive
3	***NO regularly back up to our data***	**Data manipulation to the Hospital lab sensitive X-ray dataset**	**May result in human death or serious health problems**	*Attacker will fraud the datasets so may an X-ray test result be wrong because of the inaccurate prediction due to the false dataset.*	**High**	**High**	**100**	Controls should be preventive

4	***NO Parameterized queries in the portal system***	**Sql Injection for Patient contact information**	**May noticeably affect the Portal mission, reputation or interest.**	*Attacker motivation is to know secret info. of Patient to override their valuable data, or even to execute dangerous system level commands on the database host.*	**Low**	**Low**	**1**	Controls should be detective

5	***NO Parameterized queries in the portal system***	**Sql Injection for Doctors contact information**	**May result in serious health problems**	*Attacker motivation is to know secret info. of Doctor to override their valuable data, or even to execute dangerous system level commands on the database host.*	**Medium**	**Medium**	**25**	Controls should be detective

4.1 DATASETS



Montgomery County X-ray Set	Shenzhen Hospital X-ray Set
acquired from the tuberculosis control program of the Department of Health and Human Services of Montgomery County, MD, USA. This set contains 138 posterior-anterior x-rays, of which 80 x-rays are normal and 58 x-rays are abnormal. It is a labeled dataset.	collected by Shenzhen No.3 Hospital in Shenzhen, Guangdong providence, China There are 326 normal x-rays and 336 abnormal x-rays. It is a labeled dataset with age and gender type.

Table 4: datasets used for Smart (TB) Detection

4.1.1 EXPERIMENTS OF CUSTOM CNN

The highest CNN model consists of 5 convolutional blocks, followed by a global average pooling layer (which compresses each feature map by replace it with mean value) and a fully-connected softmax layer with two outputs. Each convolutional block contains two 3 × 3 convolutions with activation function ReLUs, followed by a max-pooling. The pooling size is 3 × 3 with stride of 2, similarly to AlexNet22. Each block has implemented by a 1 × 1 convolution: the output of the shortcut is summed to that of the 3 × 3 convolutions before enter the pooling. All convolutions are containing zero-padding in order to preserve the input resolution. Each convolutional layer used the batch normalization to speed up the training procedure and reduce overfitting. The convolutions of the first block have a stride of 2 in order to reduce the amount of calculations required by the network. The resolution of the input was degraded to 512 ×

512 pixels as input. A similar strategy is used by AlexNet22. The connection of the first block has stride of 4 to match the resolution of the other convolutions. In our tests, by using the stride of two there is no effects on the accuracy even if the computational requirements significantly decreased. the first layer concerned to extract very basic features, such as edges and stripes patterns, a possible explanation for this is that the patterns can be extracted as easily with stride convolutions as with dense ones. The depth of 5 blocks because this number corresponds to receptive window that covers the whole input image, and this window size allows the network to access a large context for its decisions at each location. Empirical tests step confirmed that this depth leads to good classification performances. The input data is pre-processed with the following steps: (1) if there any black band or borders then it is cropped from the edges of each image, (2) the image is resized to be size of 512 *512 pixels (3) the central region 512 × 512 is extracted. After this, we made a calculation by using the mean over all pixels in the whole dataset is subtracted and the pixel values are divided by their standard deviation. Training was performed using categorical cross-entropy as the error function and with mini-batches of four samples. The samples are shuffled each epoch before forming the mini-batches, in order to randomize and generalize the whole learning procedure and reduce overfitting.

4.1.2 EXPERIMENTS OF ENSEMBLE MODELS

The first experiment is to use CNN vgg-16 with SVM. Vgg-16 consists of the input to convolution layer is fixed size of 224 x 224 x 3 (RGB images). The image is passed through convolution layers, where the kernel(filters) were used with a very small receptive field size of 3×3. In one of the configurations, it also utilizes 1×1 convolution filter, which can be seen as a linear transformation of the input channels (followed by non-linear). The convolution stride is set to be value of 1, the spatial padding of convolution. layer input is such that the spatial resolution is preserved after convolution, e.g. the padding is 1-pixel for 3×3 convolution layers. Vgg-16 consist of five max-pooling layers, which follow some of the convolution layers. not all convolution layers are followed by the max-pooling. Max-pooling is performed with a size of 2×2- window, with stride of 2.Three Fully-Connected layers follow a stack of convolutional layers (which has a different depth in different architectures): the first two fully connected layer have size of 4096 channels each, the third one contains contains 1000 channels (one for each class). The final layer is the softmax layer. Note that every convolution layer and fully connected layer is followed by ReLu activation function. There is no regularization (Dropout) optimization part in vgg-16. the images were classified as either TB or no-TB. In this type of model, the learning rate and epoch were set to 0.0001 and 2000 respectively.

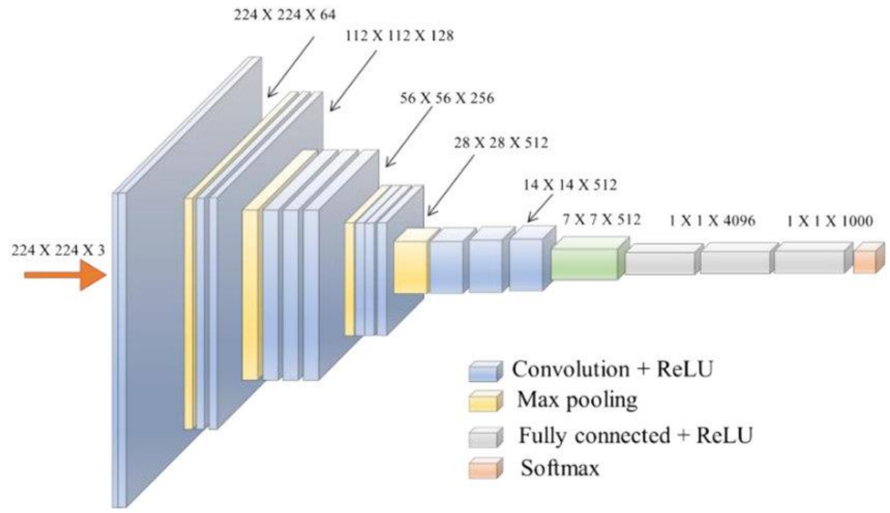


Figure 48: Vgg-16 architecture

Other experiment is done by using machine learning only hand engineer HOG feature extraction and histogram equalization then applying support vector machine as a basic classifier and random forest as ensemble classifier. The most kernel use in SVM is “rbf” with c=1 and gamma=0.01 and fed the result of support vector machine to Random forest with n_estimators=100, max_depth with value of 2 and criterion ‘’gini”. By using these steps, the achieved accuracy is 99.14%.

4.1.3 Experiment of Machine learning (Hog + SVM):

4.1.3.1 Experiment 1:

The first experiment is to use hog feature extraction with SVM. The parameter of feature extraction are orientations=4, pixels_per_cell=(16,16),cells_per_block=(2, 2). And the parameter of SVM are C=1.0, gamma='scale', kernel='rbf', tol=0.001.

4.1.3.2 Experiment 2:

The first experiment is to use hog feature extraction with SVM. The parameter of feature extraction are orientations=8, pixels_per_cell=(16,16),cells_per_block=(2, 2). And the parameter of SVM are C=1.0, gamma='scale', kernel='rbf', tol=0.001.

4.1.3.3 Experiment 3:

The first experiment is to use hog feature extraction with SVM. The parameter of feature extraction are orientations=2, pixels_per_cell=(16,16),cells_per_block=(2, 2). And the parameter of SVM are C=1.0, gamma='scale', kernel='rbf', tol=0.001.

4.1.3.4 Experiment 4:

The first experiment is to use hog feature extraction with SVM. The parameter of feature extraction are orientations=4, pixels_per_cell=(8,8),cells_per_block=(1, 1). And the parameter of SVM are C=1.0, gamma='scale', kernel='rbf', tol=0.001.

4.1.3.5 Experiment 5:

The first experiment is to use hog feature extraction with SVM. The parameter of feature extraction are orientations=4, pixels_per_cell=(16,16),cells_per_block=(1, 1). And the parameter of SVM are C=1.0, gamma='scale', kernel='rbf', tol=0.001.

4.1.3.6 Experiment 6:

The first experiment is to use hog feature extraction with SVM. The parameter of feature extraction are orientations=4, pixels_per_cell=(16,16),cells_per_block=(4, 4) . And the parameter of SVM are C=1.0, gamma='scale', kernel='rbf', tol=0.001.

4.1.3.7 Experiment 7:

The first experiment is to use hog feature extraction with SVM. The parameter of feature extraction are orientations=8, pixels_per_cell=(16,16),cells_per_block=(8, 8). And the parameter of SVM are C=1.0, gamma='scale', kernel='rbf', tol=0.001.

4.1.3.8 Experiment 8:

The first experiment is to use hog feature extraction with SVM. The parameter of feature extraction are orientations=2, pixels_per_cell=(16,16),cells_per_block=(8, 8) .And the parameter of SVM are C=1.0, gamma='scale', kernel='rbf', tol=0.001.

4.1.3.9 Experiment 9:

The first experiment is to use hog feature extraction with SVM. The parameter of feature extraction are orientations=4, pixels_per_cell=(8,8),cells_per_block=(8, 8).And the parameter of SVM are C=1.0, gamma='scale', kernel='rbf', tol=0.001.

4.1.3.10 Experiment 10:

The first experiment is to use hog feature extraction with SVM. The parameter of feature extraction are orientations=4, pixels_per_cell=(16,16),cells_per_block=(8, 8).And the parameter of SVM are C=1.0, gamma='scale', kernel='rbf', tol=0.001.

4.1.3.11 Best Experiment result:

Experiment hog feature extraction with principle component analysis (PCA) and classification using SVM, it gets the highest accuracy in the machine learning approach, the experiment contains these parameters are orientations=4, pixels_per_cell=(16,16), cells_per_block=(2, 2) and the n_components in PCA equal to 800. And the parameters of SVM classifier are C equal to 1.0 gamma equal to scale kernel equal to rpf and tol equal to 0.001

4.1.4 RESULTS

Convolution neural network:

CNN using Keras implementation of InceptionV3, VGG16,AlexNet,LeNet and custom model architectures. these architectures were pretrained using the ImageNet x-ray dataset except our custom model. datasets consist of Montgomery and Shenzhen datasets and consist of 800 images.

CNN Architecture	Accuracy (%)
VGG16	69.7%
InceptionV3	67.4%
AlexNet	64.08%
Custom CNN	62.773%
LeNet	65.9%

Table 5: Performance of Different single CNN Architectures to TB Detection.

CNN feature extractor	Feature selector	Classification algorithm	precision (%)	recall (%)	F1-score (%)	Support(%)	Accuracy(%)
Vgg-16	PCA	Svm	90	90	90	80	90

Table 6: Our main project model that gets the highest accuracy.

Experiment	Accuracy
Experiment 1	82.5%
Experiment 2	81.875%
Experiment 3	81.875%
Experiment 4	82.5%
Experiment 5	80.625%
Experiment 6	82.5%
Experiment 7	83.75%
Experiment 8	82.5%
Experiment 9	84.375%
Experiment 10	85.625%

Table 7: Experiments.

4.2 ANALYSIS/DISCUSSION

Deep learning approach especially CNN plays a vital algorithm in tuberculosis prediction, it used as a classification algorithm even the model is a custom model or pre-trained model. Additionally, CNN not only used as classification algorithm it can be used as feature extractor with the accurate result, but in the tuberculosis disease results, CNN achieves good results but it is not the highest accuracy and there are other models can compete with CNN for achieving more accurate results e.g. ensemble learning (e.g. SVM +RF). When the dataset is large enough CNN can achieve the highest accuracy between all models. Ensemble learning proves in this study it is a robust learning approach and can achieve the highest accuracy and solve other problems not only tuberculosis disease. In future work, if we can collect more datasets that will be good enough to generalize any type of model and achieves a high accuracy especially deep learning approach. Provide an enhancing technique, in contrast, histogram equalization and feature selection (e.g. PCA – principal component analysis).

Learning approach	Advantages	Disadvantages
Deep learning	Achieves high accuracy when the data is large. Provide Automatic feature extraction. CNNs can work as a classifier or as feature extractor. CNNs provide custom model or transfer learning.	Requires large amount of data for working accurately. Need to have a strong GPU computational capability. CNN is black box model like ANN we can't estimate what the relation between layers or how it controls the network.
Ensemble learning	Achieves high Accuracy.	When data is more complex it is hard to

	Ensemble used multiple algorithms for enhancing the accuracy or reduce the error from last classifier. It is a robust an approach in predicting medical imaging especially tuberculosis disease. No need to have a large or big dataset like deep learning.	generate a generalized model. Need to make a hand engineer feature extraction unlike deep learning it provides an automatic feature extraction.

Table 8: Advantages and disadvantages in deep and ensemble learning.

This work shows compelling results, and provides a good foundation for further work. This work provides a review of various image processing methods followed for the detection of M. tuberculosis from chest x-ray images. There are four main steps in a Smart (TB) system: collecting chest x-rays, apply an appropriate pre-processing steps, extracting important features, and classifying disease according to the features. In data pre-processing and extraction of features, the techniques of enhancement and segmentation are very important. Usually, there are many ways to highlight lesions and suppress noise. In the segmentation, the deformable model and the deep learning method are the best, while the rule-based methods have poor performance, and they often used together with other methods to improve the segmentation performance. The real challenge that lies ahead is determining how such DL systems will realistically fit into the workflow of specific clinical screening or diagnostic infective disease settings. CNN model is the most powerful algorithm for detecting TB in Smart (TB) and other related works .Artificial intelligence can facilitate diagnosis TB in its early stages. Smart TB Portal will provide many features that can facilitate the communication between different types of users, facilitate booking for medical laboratories and chest clinics.

5.1 FUTURE WORK

In a Smart (TB) we are focusing on how the model will work correctly and suitable for the pervious requirement which includes accuracy, speed scientific and practice efficiency. We can enhance these parts in future by using:

1- Get more data of chest x-ray because the more images with efficient pre-processing in deep learning the better model and prediction.
2- Provide another type of data for TB disease to make sure that the decision of chest x-ray model matches the result for the second model.

In a Smart (TB) portal we can provide more features that can help to attract more users and facilitate the communication between all users for example:

1- Provide promo-codes that patients can use it for booking appointment in chest x-ray laboratory center and chest clinics.
2- Booking for medical canters or hospitals by therapist doctor if patient condition is very serious.
3- Develop new category of users which is the Hospital Doctors with its features that is concerned to x-ray samples diagnosis.
4- Change the database from local host database to cloud database.
5- Develop new feature for patients which is answering on symptoms questions and write medical complains.
6- Develop new feature for doctors which is reviewing on symptoms questions and write medical complains.

REFERENCES

[1] Ensemble deep learning for tuberculosis detection using chest X-ray and canny edge detected images M Hijazi, S Kieu Tao Hwa, A Bade, R Yaakob, M Saffree Jeffree, 2019.

[2] Norval, M., Wang, Z. and Sun, Y., 2019. Pulmonary Tuberculosis Detection Using Deep Learning Convolutional Neural Networks. *Proceedings of the 3rd International Conference on Video and Image Processing,*

[3] Norval, M., Wang, Z. and Sun, Y., 2019. Pulmonary Tuberculosis Detection Using Deep Learning Convolutional Neural Networks. *Proceedings of the 3rd International Conference on Video and Image Processing,*

[4] Rajaraman, S. and Antani, S., 2020. Modality-Specific Deep Learning Model Ensembles Toward Improving TB Detection in Chest Radiographs. *IEEE Access*, 8, pp.27318-27326.

[5] Asha.T, S.Natarajan, K. N. B. Murthy.,2011. Effective-Classification-Algorithms-to-Predict-the-Accuracy-of-Tuberculosis-A-Machine-Learning-approach International Journal of Computer Science And Information Technology,9.

[6] Pasa, F., Golkov, V., Pfeiffer, F. *et al.* Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. *Sci Rep* 9**,** 6268 (2019),10.1038/s41598-019-42557-4

[7] O. Yadav, K. Passi and C. K. Jain, "Using Deep Learning to Classify X-ray Images of Potential Tuberculosis Patients," *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Madrid, Spain, 2018, pp. 2368-2375, doi: 10.1109/BIBM.2018.8621525.

[8] https://web.stanford.edu/class/cs231a/lectures/intro_cnn

[9] Daniel Shu Wei Ting, Tien-En Tan, C C Tchoyoson Lim, Development and Validation of a Deep Learning System for Detection of Active Pulmonary Tuberculosis on Chest Radiographs: Clinical and Technical Considerations, *Clinical Infectious Diseases*, Volume 69, Issue 5, 1 September 2019, Pages 748–750,

[10] https://apps.who.int/iris/bitstream/handle/10665/329368/9789241565714-eng.pdf?ua=1

[11] D. Varshni, K. Thakral, L. Agarwal, R. Nijhawan and A. Mittal, "Pneumonia Detection Using CNN based Feature Extraction," *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, India, 2019, pp. 1-7, doi: 10.1109/ICECCT.2019.8869364.

[12] Lakhani, P., & Sundaram, B. (2017). Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology*, *284*(2), 574-582. doi: 10.1148/radiol.2017162326

[13] Detection of Tuberculosis in Chest X-rays using U-Net Architecture. (2019). *International Journal Of Innovative Technology And Exploring Engineering*, *9*(1), 2514-2519. doi: 10.35940/ijitee.a4834.119119

[14] Detection of Pulmonary Tuberculosis Manifestation in Chest X-Rays using Different Convolutional Neural Network (CNN) Models. (2019). *International Journal Of Engineering And Advanced Technology*, *9*(1), 2270-2275. doi: 10.35940/ijeat.a2632.109119

[15] El-Solh, A., Hsiao, C., Goodnough, S., Serghani, J., & Grant, B. (1999). Predicting Active Pulmonary Tuberculosis Using an Artificial Neural Network. *Chest*, *116*(4), 968-973. doi: 10.1378/chest.116.4.968

[16] Rahman, T., Chowdhury, M., Khandakar, A., Islam, K., Islam, K., & Mahbub, Z. et al. (2020). Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using Chest X-ray. *Applied Sciences*, *10*(9), 3233. doi: 10.3390/app10093233
[17] Pongpirul, K., Sathitratanacheewin, S., & Sunanta, P. (2018). 1992. Automated Classification of Pulmonary Tuberculosis-Associated Radiograph in the US Hospital-Scale Chest X-ray Database by Using Deep Convolutional Neural Network. *Open Forum Infectious Diseases*, *5*(suppl_1), S579-S579. doi: 10.1093/ofid/ofy210.1648

[18] Y. Li and G. Su, "Simplified histograms of oriented gradient features extraction algorithm for the hardware implementation," *2015 International Conference on Computers, Communications, and Systems (ICCCS)*, Kanyakumari, 2015, pp. 192-195, doi: 10.1109/CCOMS.2015.7562899.