AI-DocHelper

Submitted in partial fulfillment of the requirements
of the degree of

BACHELOR OF ENGINEERING

In

COMPUTER ENGINEERING

By
Group No: 43

Roll No.	Name
1704130	Anuja Kothavale
1704137	Vikrant Shah
1704140	Surbhi Singh

Guide:

PROF. Aejaz khan
(Assistant Professor, Department of Computer Engineering, TSEC)

THADOMAL SHAHANI
TSEC
ENGINEERING COLLEGE

Computer Engineering Department
Thadomal Shahani Engineering College
University of Mumbai
2020-2021

CERTIFICATE

This is to certify that the project entitled **“AI-DocHelper”** is a bonafide work of

Roll No.	Name
1704130	Anuja Kothavale
1704137	Vikrant Shah
1704140	Surbhi Singh

Submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **“BACHELOR OF ENGINEERING”**in**“COMPUTER ENGINEERING”**.

Prof. Aejaz Khan
Guide

Dr. Tanuja Sarode
Head of Department

Dr.G.T.Thampi
Principal

Project Report Approval for B.E

Project report entitled (***AI-DocHelper***) by

Roll No.	Name
1704130	Anuja Kothavale
1704137	Vikrant Shah
1704140	Surbhi Singh

is approvedfor the degree of ***"BACHELOR OF ENGINEERING" in "COMPUTER ENGINEERING"***.

Examiners

1.________________________________

2.________________________________

Date:

Place:

Declaration

We declare that this written submission represents my ideas in my own words and where others 'ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have a adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will because for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

1) ____________________

Anuja Kothavale, 1704130

2) ____________________

Vikrant Shah, 1704137

3) ____________________

Surbhi Singh, 1704140

Date:

Abstract

For many years, many people have died due to undetected diseases. Early detection of these diseases at the micro classification stage can be useful for providing proper treatment to the patients at the early stage and could have saved a lot of lives. A lot of research is being done to detect these diseases at the earliest. Therefore, a computer-aided or artificial intelligence approach for detecting diseases at the early stage is being proposed, which makes use of machine learning and deep learning algorithms for detecting diseases. This system will detect all general diseases such as different types of cancer, malaria, diabetic retinopathy, etc. AI-DocHelper is being proposed as there is no system available, that detects all these general diseases.

AI-DocHelper system will be able to detect diseases at a very early stage. It will take all the factors responsible or an image (to get the presence of virus, tumor, etc.) of a patient for detecting the particular disease. It will also store all the results in the system for the future use.

TABLE OF CONTENTS

Sr. No.	Topic	Page No.
	List of Figures	i
	List of Tables	ii
1.	**Introduction**	1
	1.1 Introduction	1
	1.2 Aims & Objective	2
	1.3 Scope	2
2.	**Review of Literature**	3
	2.1 Domain Explanation	3
	2.1.1 Machine Learning	3
	2.1.2 Deep Learning	5
	2.2 Hardware & Software Requirement	6
	2.2.1 Hardware Requirements	6
	2.2.2 Software Requirements	6
3.	**Analysis**	8
	3.1 Functional Requirement	8
	3.2 Non-Functional Requirement	8
	3.3 Proposed System	9
	3.3.1 Dataset	11
4.	**Design**	20
	4.1 Design Consideration	20
	4.2 Design Details	22
	4.2.1 Class Diagran	22

4.2.2 Sequence Diagram 22

4.2.3 Data Preprocessing 23

4.2.4 Comparative Study for Image Dataset 25

4.2.5 Study for Image Dataset 27

4.2.6 Algorithm 27

4.3 GUI Design 30

5. Implemantation 35

5.1 Plan of Implementation 35

5.2 Results And Evaluation 36

6. **Comclusion** 42

Appemdix

References 43

Acknowledgement 45

List of Figures

Figure No.	Description	Page No.
Figure 3.1	Functional Requirements	8
Figure 3.2	Block Diagram for Training the Model	10
Figure 3.3	Applying Real Time Data to the Model	11
Figure 4.1	System Design Workflow for Textual Dataset	20
Figure 4.2	System Design Workflow for Image Dataset	21
Figure 4.3	Class diagram for Real Time Data	22
Figure 4.4	Sequence diagram for Real Time Data	23
Figure 4.5	Confusion Matrix for Diabities Detector with Accuracy	26
Figure 4.6	Random Forest Classifier	28
Figure 4.7	Convolutional Nueral Network	29
Figure 4.8	Sign Up Page	30
Figure 4.9	Sign In Page	31
Figure 4.10	Forgot Password Page	31
Figure 4.11	Home Page	32
Figure 4.12	Input Page For Image Dataset	33
Figure 4.13	Result Page for Image Dataset	33
Figure 4.14	Input Page for Textual Dataset	34
Figure 5.1	Accuracy Graph and Loss Graph of Models	37

List of Tables

Table No.	Description	Page No.
Table 4.1	Comparison of Accuracy for Textual Dataset	26
Table 5.1	Plan of Implementation	35
Table 5.2	Accuracy & Loss of All Models	36

Chapter 1

Introduction

1.1 Introduction

Healthcare is one of the most urgent matters in human societies, as the life quality of citizens directly depends on it [1]. However, the healthcare sector is highly heterogeneous, widely distributed and fragmented. From the clinical perspective, delivering appropriate patient care requires access to relevant patient information, which is seldom available where and when it is needed [2]. Additionally, the wide variation in test-ordering for diagnostic purposes suggests the requirement of sufficient and appropriate test set [3,4]. Smellie et al. [5] extended this argument by suggesting that the large differences observed in general practice pathology requesting result mostly from individual variation in clinical practice and are, therefore, potentially susceptible to change through more consistent and better informed decision-making for doctors [6].

Hence, medical data often consist of a large set of heterogeneous variables, collected from different sources, such as demographics, disease history, medication, allergies, biomarkers, medical images, or genetic markers, each of which offer a different partial view on a patient's state. Moreover, statistical properties among the aforementioned sources are inherently different. When researchers and practitioners analyse such data, they are confronted with two problems: the curse of dimensionality (the feature space is increasing exponentially in the number of dimensions and the number of samples), and the heterogeneity in feature sources and statistical properties [7]. These factors provoke delays and inaccuracy in the disease detection and, consequently, patients could not receive the appropriate cares [8]. Thus, there is a clear need for an effective and robust methodology that allows for the early disease detection and it can be used by doctors as a help for decision-making [9]. Therefore, medical, computational, and statistical fields are facing the challenge of exploring new techniques for modeling the prognosis and diagnosis of diseases, since traditional paradigms fail in the treatment of all this information [10]. This requirement is quite related to the evolutions in other domains, such as Big Data (BD), Data Mining (DM), or Artificial Intelligence (AI).

1.2 Aim & Objectives

The main aim is to design a system, that can detect or identify all basic diseases (such as diabetes, etc.) as well as all the deadly diseases (such as cancer, tumor, etc.) at a micro classification stage with the help of some tests. Once the disease is detected at an early stage, it can be cured with the help of proper medical procedures and can even save the patient’s life.

The objective of this project is as follows :

- To develop models that can detect diseases.
- Store user data for future use.
- Give/Generate precise results.
- Provide a user-friendly GUI.

1.3 Scope

- The system will be a web-based system that provides early detection of diseases using some suitable algorithm.
- A user-friendly GUI is provided which makes it convenient for the end-users.
- Stores user’s test results. So, it can be used in the future for reference.
- The system will not provide a cure for the disease.
- After collecting the sample for the test, if the user gets a disease, the system will not be able to detect it.

Chapter 2

Review of Literature

2.1 Domain Explanation

2.1.1 Machine Learning

- **Logistic Regression –**

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

- **KNN Algorithm –**

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

- **SVM Classifier –**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

- **Naïve Bayes Classifier –**

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in *text classification* that includes a high-dimensional training dataset.

Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

- **Decision Tree Classifier –**

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset. *It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.* It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

- **Random Forest Algorithm –**

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.*

As the name suggests, *"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."* Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

2.1.2 Deep Learning :

- **Artificial Neural Network –**

The term "Artificial neural network" refers to a biologically inspired sub-field of artificial intelligence 14odelled after the brain. An Artificial neural network is usually a computational network based on biological neural networks that construct the structure of the human brain. Similar to a human brain has neurons interconnected to each other, artificial neural networks also have neurons that are linked to each other in various layers of the networks. These neurons are known as nodes.

Artificial neural network tutorial covers all the aspects related to the artificial neural network. In this tutorial, we will discuss ANNs, Adaptive resonance theory, Kohonen self-organizing map, Building blocks, unsupervised learning, Genetic algorithm, etc.

- **Convolutional Neural Network –**

Convolutional Neural Network is one of the main categories to do image classification and image recognition in neural networks. Scene 15abelling, objects detections, and face recognition, etc., are some of the areas where convolutional neural networks are widely used.

CNN takes an image as input, which is classified and process under a certain category such as dog, cat, lion, tiger, etc. The computer sees an image as an array of pixels and depends on the resolution of the image. Based on image resolution, it will see as h * w * d, where h= height w= width and d= dimension. For example, An RGB image is 6 * 6 * 3 array of the matrix, and the grayscale image is 4 * 4 * 1 array of the matrix.

In CNN, each input image will pass through a sequence of convolution layers along with pooling, fully connected layers, filters (Also known as kernels). After that, we will apply the Soft-max function to classify an object with probabilistic values 0 and 1.

2.2 Hardware & Software Requirements

2.2.1 Hardware Requirements

- Processor : Intel Core i3 / Pentium
- RAM : 4 GB Minimum
- Hard Disk Space : 30 GB Minimum

2.2.2 Software Requirements

- Operating System : Windows / MacOS / Linux
- Browser : Google Chrome / Mozilla Firefox
- Python 3.x
- Tensorflow 2.0
- Annaconda
- Jupyter Notebook
- Flask
- Numpy

- Pandas
- Scikit learn
- HTML5
- Javascript
- CSS3
- Bootstrap 5.0
- SQL_Alchemy

Chapter 3

Analysis

3.1 Functional Requirements

A **Functional Requirement** (FR) is a description of the service that the software must offer. It describes a software system or its component. A function is nothing but inputs to the software system, its behavior, and outputs. It can be a calculation, data manipulation, business process, user interaction, or any other specific functionality which defines what function a system is likely to perform. Functional Requirements are also called Functional Specification.

In software engineering and systems engineering, a Functional Requirement can range from the high-level abstract statement of the sender's necessity to detailed mathematical functional requirement specifications. Functional software requirements help you to capture the intended behavior of the system.

The functional requirements of this system are :

- Early detection of disease : Upon entering certain symptoms, this system can name particular disease that has those symptoms.
- User friendly GUI : User Interface is simple and easy to use.
- Precise results : System provides the most accurate result for the particular symptom.
- User Data Store : General information and recent data of the user will be displayed

Input
System Behaviour
Output

Fig 3.1: Functional Requirements

3.2 Non-Functional Requirements

NON-FUNCTIONAL REQUIREMENT (NFR) specifies the quality attribute of a software system. They judge the software system based on Responsiveness, Usability, Security, Portability, and other non-functional standards that are critical to the success of the software system. An example of a nonfunctional requirement, *“how fast does the website load?”* Failing to meet non-functional requirements can result in systems that fail to satisfy user needs.

Non-functional Requirements allows you to impose constraints or restrictions on the design of the system across the various agile backlogs. Example, the site should load in 3 seconds when the number of simultaneous users are > 10000. Description of non-functional requirements is just as critical as a functional requirement.

The non-function of this system are as follows :

- Usability : Easy to use.
- Reliability : Product performs in the same way as it was expected to.
- Performance : Performance is good.
- Secure : User data is safe.

3.3 Proposed System

AI-DocHelper, an artificial intelligence-based system that can detect general diseases at an early stage, will take textual and/or image data as input. All relevant features will be extracted from the input data. For a disease to be detected, different classification models will be applied to the feature extracted training data. The model that gives the best accuracy and minimum loss in the testing phase will be selected as the final model for detecting a particular disease. The output or final result is the classification of whether a particular patient has a particular disease or not

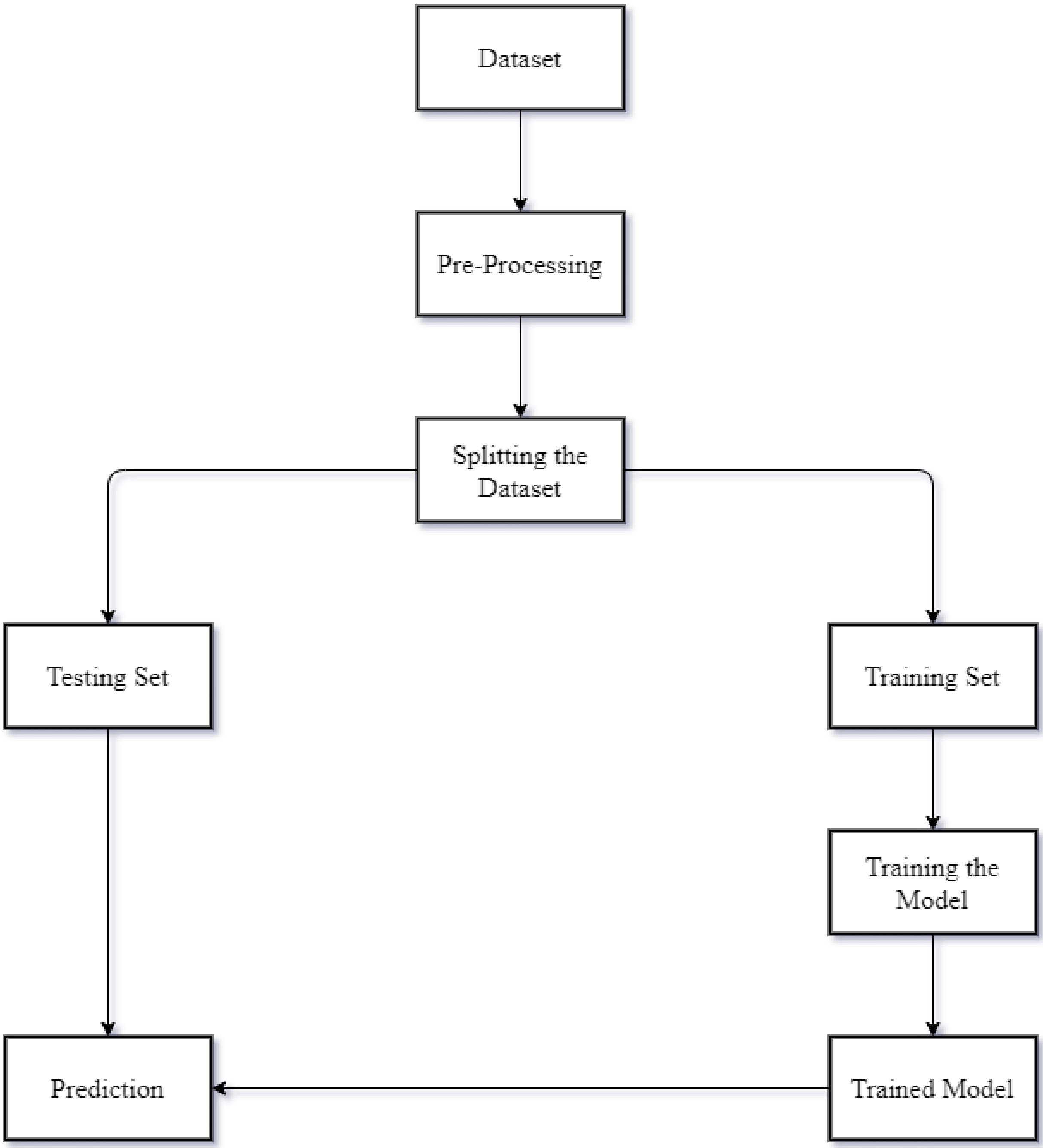


Fig 3.2: Block Diagram for Training the Model

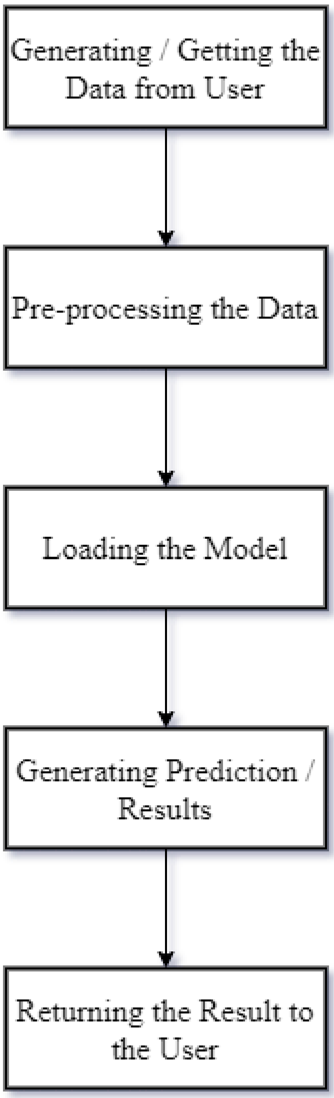

Generating / Getting the Data from User
Pre-processing the Data
Loading the Model
Generating Prediction / Results
Returning the Result to the User

Fig 3.3: Applying Real Time Data to the Model

3.3.1 Dataset

3.3.1.1 Brain Tumor Classification (MRI) Dataset

A Brain tumor is considered as one of the aggressive diseases, among children and adults. Brain tumors account for 85 to 90 percent of all primary Central Nervous System(CNS) tumors. Every year, around 11,700 people are diagnosed with a brain tumor. The 5-year survival rate for people with a cancerous brain or CNS tumor is approximately 34 percent for men and 36 percent for women. Brain Tumors are classified as Benign Tumor, Malignant Tumor, Pituitary Tumor, etc. Proper treatment, planning, and accurate diagnostics should be implemented to improve the life expectancy of the patients. The best technique to detect brain tumors is Magnetic Resonance Imaging (MRI). A huge amount of image data is generated through the scans. These images are examined by the radiologist. A manual examination can be error-prone due to the level of complexities involved in brain tumors and their properties.

Dataset Link : https://www.kaggle.com/sartajbhuvaji/brain-tumor-classification-mri
No. of Samples : 3264
Classes : No Tumor, Glioma Tumor, Meningioma Tumor, Pituitary Tumor

- **Glioma Tumor**

A type of tumor that occurs in the brain and spinal cord. Gliomas can occur in the brain and various locations in the nervous system, including the brain stem and spinal column. Different types of gliomas cause different symptoms. Some include headaches, seizures, irritability, vomiting, visual difficulties, and weakness or numbness of the extremities. Treatments include surgery, radiation therapy, chemotherapy, and targeted molecular therapy.

- **Meningioma Tumor**

A usually non-cancerous tumor that arises from the membranes surrounding the brain and spinal cord. It isn't clear what causes a meningioma. Radiation therapy, female hormones, and genetics may play a role. In most cases, the condition is non-cancerous. Symptoms depend on the size of the tumor, changes in vision, headaches, hearing loss, and seizures. A small, slow-growing meningioma that isn't causing signs or symptoms may not require treatment. When required, treatment might involve surgery or radiation.

- **Pituitary Tumor**

Non-cancerous tumors in the pituitary gland don't spread beyond the skull. The pituitary gland is in the skull, below the brain, and above the nasal passages. A large tumor can press upon and damage the brain and nerves. Vision changes or headaches are symptoms. In some cases, hormones can also be affected, interfering with menstrual cycles and causing sexual dysfunction. Treatments include surgery and medication to block excess hormone production or shrink the tumor. In some cases, radiation may also be used.

3.3.1.2 Breast Cancer Dataset

Cancer that forms in the cells of the breasts. Breast cancer can occur in women and rarely in men. Symptoms of breast cancer include a lump in the breast, bloody discharge from the nipple, and changes in the shape or texture of the nipple or breast. Its treatment depends on the stage of cancer. It may consist of chemotherapy, radiation, hormone therapy, and surgery.

Dataset Link : https://scholar.cu.edu.eg/?q=afahmy/pages/dataset

No. of Samples : 1578 samples

Classes : Normal, Benign, Malignant

3.3.1.3 Colon Cancer Dataset

Cancer of the colon or rectum, located at the digestive tract's lower end. Early cases can begin as non-cancerous polyps. These often have no symptoms but can be detected by screening. For this reason, doctors recommend screenings for those at high risk or over the age of 50. Colorectal cancer symptoms depend on the size and location of cancer. Some commonly experienced symptoms include changes in bowel habits, changes in stool consistency, blood in the stool, and abdominal discomfort. Colorectal cancer treatment depends on the size, location, and how far cancer has spread. Common treatments include surgery to remove cancer, chemotherapy, and radiation therapy.

Dataset Link : https://datasets.simula.no/kvasir/

No. of Samples : 4000 images

Classes : dyed-lifted-polyps, dyed-resection-margins, esophagitis, normal-cecum, normal-pylorus, normal-z-line, polyps, ulcerative-colitis

3.3.3.4 Pneumonia X-ray Images Dataset

Infection that inflames air sacs in one or both lungs, which may fill with fluid. With pneumonia, the air sacs may fill with fluid or pus. The infection can be life-threatening to anyone, but particularly to infants, children, and people over 65. Symptoms include a cough with phlegm or pus, fever, chills, and difficulty breathing. Antibiotics can treat many forms of pneumonia. Some forms of pneumonia can be prevented by vaccines.

Dataset Link : https://www.kaggle.com/pcbreviglieri/pneumonia-xray-images

No. of Samples : 5856 images

Classes : normal, opacity

3.3.3.5 Covid-19 Radiology Database

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Most people who fall sick with COVID-19 will experience mild to moderate symptoms and recover without special treatment. The virus that causes COVID-19 is mainly transmitted through droplets generated when an infected person coughs, sneezes, or exhales. These droplets are too heavy to hang in the air and quickly fall on floors or surfaces. You can be infected by breathing in the virus if you are within proximity of someone who has COVID-19, or by touching a contaminated surface and then your eyes, nose, or mouth.

Dataset Link : https://www.kaggle.com/tawsifurrahman/covid19-radiography-database

No. of Samples : 21,000 images

Classes : normal, lung opacity, COVID

3.3.3.6 Leukemia Classification

A type of cancer of the blood and bone marrow that affects white blood cells. Acute lymphoblastic leukemia is the most common childhood cancer. It occurs when a bone marrow cell develops errors in its DNA. Symptoms may include enlarged lymph nodes, bruising, fever, bone pain, bleeding from the gums, and frequent infections. Treatments may include chemotherapy or targeted drugs that specifically kill cancer cells.

ALL is the most common type of childhood cancer and accounts for approximately 25% of pediatric cancers. These cells have been segmented from microscopic images and are representative of images in the real world because they contain some staining noise and illumination errors, although these errors have largely been fixed in the course of acquisition. The task of identifying immature leukemic blasts from normal cells under the microscope is challenging due to morphological similarity and thus the ground truth labels were annotated by an expert oncologist.

Dataset Link : https://www.kaggle.com/andrewmvd/leukemia-classification

No. of Samples : 3527 images

Classes : all, hem

3.3.3.7 Breast Histalopathy Images Dataset

Invasive ductal carcinoma (IDC), sometimes called infiltrating ductal carcinoma, is the most common type of breast cancer. About 80% of all breast cancers are invasive ductal carcinomas. *Invasive* means that cancer has "invaded" or spread to the surrounding breast tissues. *Ductal* means that cancer began in the milk ducts, which are the "pipes" that carry milk from the milk-producing lobules to the nipple. *Carcinoma* refers to any cancer that begins in the skin or other tissues that cover internal organs — such as breast tissue. All together, "invasive ductal carcinoma" refers to cancer that has broken through the wall of the milk duct and begun to invade the tissues of the breast. Over time, invasive ductal carcinoma can spread to the lymph nodes and possibly to other areas of the body

Dataset Link : https://www.kaggle.com/paultimothymooney/breast-histopathology-images

No. of Samples : 2000 images

Classes : 0, 1

3.3.3.8 Heart Disease UCI Dataset

Dataset Link : https://www.kaggle.com/ronitf/heart-disease-uci

Attribute Information :

- age
- sex
- chest pain type (4 values)

- resting blood pressure
- serum cholestoral in mg/dl
- fasting blood sugar > 120 mg/dl
- resting electrocardiographic results (values 0,1,2)
- maximum heart rate achieved
- exercise induced angina
- oldpeak = ST depression induced by exercise relative to rest
- the slope of the peak exercise ST segment
- number of major vessels (0-3) colored by flourosopy
- thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

3.3.3.9 Chronic Kidney Disease Dataset

Our kidneys perform an important function to help filter blood and pass waste as urine. Chronic kidney disease, also called chronic kidney failure, describes the gradual loss of this function. At advanced stages, dangerous levels of fluid, electrolytes and wastes can build up in the body. Once this happens, patients must go through dialysis or consider a transplant. Our goal in this project is to see if we can predict if a patient will have chronic kidney disease or not using 24 predictors. If we are able to find variables with a strong influence on kidney failure, we may be able to detect and help patients at risk to prevent it.

Dataset Link : https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease

Attribute Information :

- age: age in years
- bp: Blood pressure in mm of Hg.
- sg: Specific Gravity
- al: Albumin - (0,1,2,3,4,5)
- su: Sugar - (0,1,2,3,4,5)
- rbc: Red Blood Cells - (normal,abnormal)
- pc: Pus Cell - (normal,abnormal)
- pcc: Pus Cell clumps - (present,notpresent)
- ba: Bacteria - (present,notpresent)
- bgr: Blood Glucose Random(numerical) in mgs/dl
- bu: Blood Urea in mgs/dl
- sc: Serum Creatinine in mgs/dl

- sod: Sodium in mEq/L
- pot: Potassium in mEq/L
- hemo: Hemoglobin in gms
- pcv: Packed Cell Volume
- wbcc: White Blood Cell Count in cells/cumm
- rbcc: Red Blood Cell Count in millions/cmm
- htn: Hypertension - (yes,no)
- dm: Diabetes Mellitus - (yes,no)
- cad: Coronary Artery Disease - (yes,no)
- appet: Appetite - (good,poor)
- pe: Pedal Edema - (yes,no)
- ane: Anemia - (yes,no)

3.3.3.10 Diabetes Disease Detection Dataset

Diabetes mellitus, commonly known as diabetes, is a metabolic disease that causes high blood sugar. The hormone insulin moves sugar from the blood into your cells to be stored or used for energy. With diabetes, your body either doesn’t make enough insulin or can’t effectively use the insulin it does make. Untreated high blood sugar from diabetes can damage your nerves, eyes, kidneys, and other organs.

Dataset Link : https://www.kaggle.com/uciml/pima-indians-diabetes-database

Attribute Information :

- Pregnancies : Number of times pregnant
- Glucose : Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure : Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skinfold thickness (mm)
- Insulin : 2-Hour serum insulin (mu U/ml)
- BMI : Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction : Diabetes pedigree function (It gives the history of diabetes in the family)
- Age : Age (years)
- Outcome : Class variable (0 or 1)

3.3.3.11 Skin Cancer: Malignant vs Benign Dataset

The most serious type of skin cancer. Melanoma occurs when the pigment-producing cells that give color to the skin become cancerous. Symptoms might include a new, unusual growth or a change in an existing mole. Melanomas can occur anywhere on the body. Treatment may involve surgery, radiation, medication, or in some cases, chemotherapy.

Dataset Link : https://www.kaggle.com/fanconic/skin-cancer-malignant-vs-benign

No. of Samples : 3297 images

Classes : benign, malignant

3.3.3.12 Malaria Cell Image Dataset

A disease caused by a plasmodium parasite, transmitted by the bite of infected mosquitoes. The severity of malaria varies based on the species of plasmodium. Symptoms are chills, fever, and sweating, usually occurring a few weeks after being bitten. People traveling to areas where malaria is common typically take protective drugs before, during, and after their trip. Treatment includes antimalarial drugs.

Dataset Link : https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria

No. of Samples : 27.6k Samples

Classes : Uninfected, Parasitized

3.3.3.13 Tuberculosis (TB) Chest X-ray Database

Tuberculosis (TB) is an infectious disease usually caused by Mycobacterium tuberculosis (MTB) bacteria. Tuberculosis generally affects the lungs, but can also affect other parts of the body. Most infections show no symptoms, in which case it is known as latent tuberculosis. A potentially serious infectious bacterial disease that mainly affects the lungs.

The bacteria that cause TB are spread when an infected person coughs or sneezes. Most people infected with the bacteria that cause tuberculosis don't have symptoms. When symptoms do occur, they usually include a cough (sometimes blood-tinged), weight loss, night sweats and fever. Treatment isn't always required for those without symptoms. Patients with active symptoms will require a long course of treatment involving multiple antibiotics.

Dataset Link : https://www.kaggle.com/tawsifurrahman/tuberculosis-tb-chest-xray-dataset?select=TB_Chest_Radiography_Database

No. of Samples : 7000 Samples

Classes : Normal, Tuberculosis

Chapter 4

Design

4.1 Design Consideration

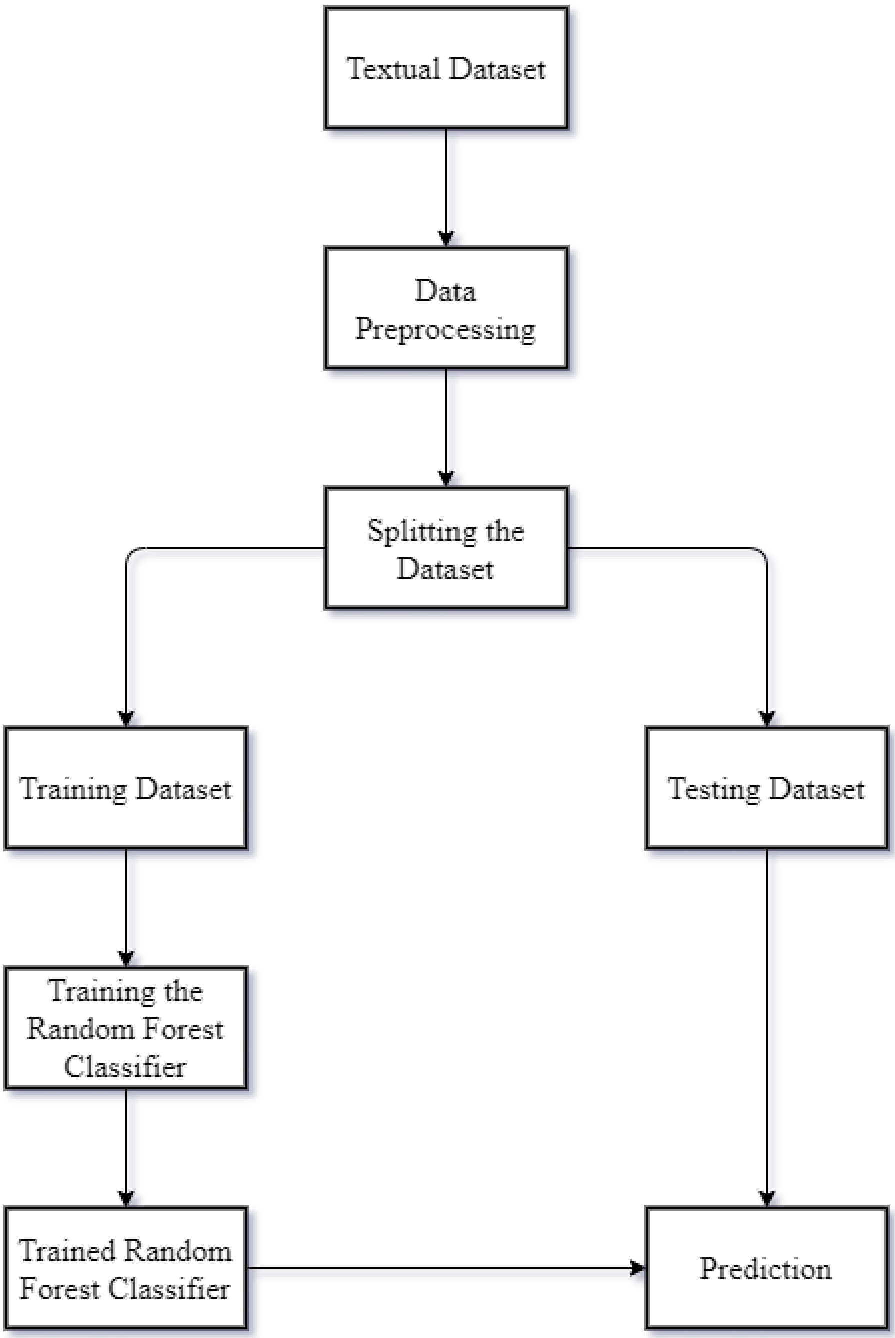


Fig 4.1: System Design Workflow for Textual Dataset

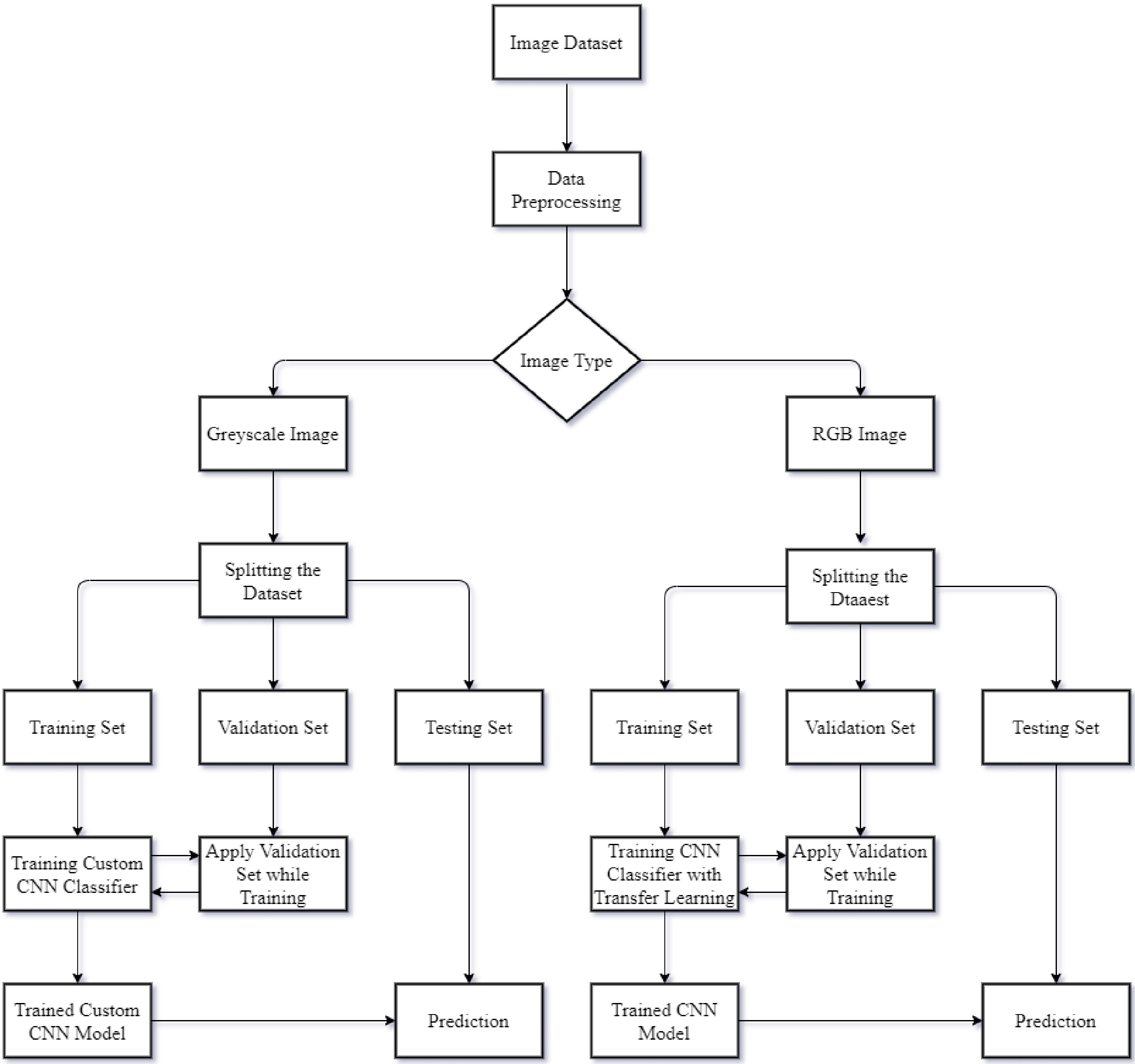


Fig 4.2: System Design Workflow for Image Dataset

Factors considered are as follows :

- Field missing in the dataset : The fields that are missing in the dataset will be replaced by NaN.
- Removing the irrelevant fields : All the irreverent files will be removed or dropped, which will not be used in training the model (such as name, phone number, etc.).
- Conversion of textual fields : Fields with textual data will be conerted to an number.
- Removing the blur images : All the blur imageswill be dropped.

4.2 Design Details

4.2.1 Class Diagram

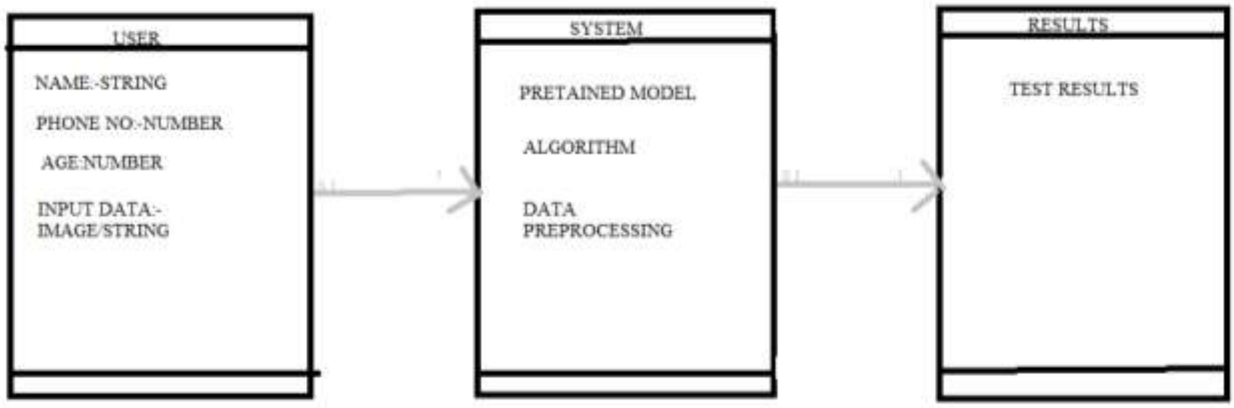


Fig 4.3: Class diagram for real-Time Data

The class diagram is shown in fig. 4.3 is for the time when the real-time data is applied to the model. It is as follows :

- User : The user data will be fed into the data. It contains personal details and all the fields for the disease to be detected.
- System : All the user data is fed to the system. The model is selected for which the disease is to be detected. The model is loaded. All the real-time data is applied to the model.
- Results : The model predicts the results for the disease and the results are shown on the screen.

4.2.2 Sequence Diagram

The sequence diagram is shown in fig. 4.4 for the project. It shows the flow events once the system is implemented and running in real-time.

The patient submits/gives all the required data (tissue samples, X-ray, CT scan, etc.). The pathologist will capture a high-quality image or record the observations. All the recorded data is fed into the system. Once the data is uploaded to the server, some preprocessing is performed on the data. Once the preprocessing is completed, the required disease predictor model is loaded, and the data is fed to the model. The model detects the disease and gives the result back to the pathologist, in the report format. Once the report is generated, it is given to the patient.

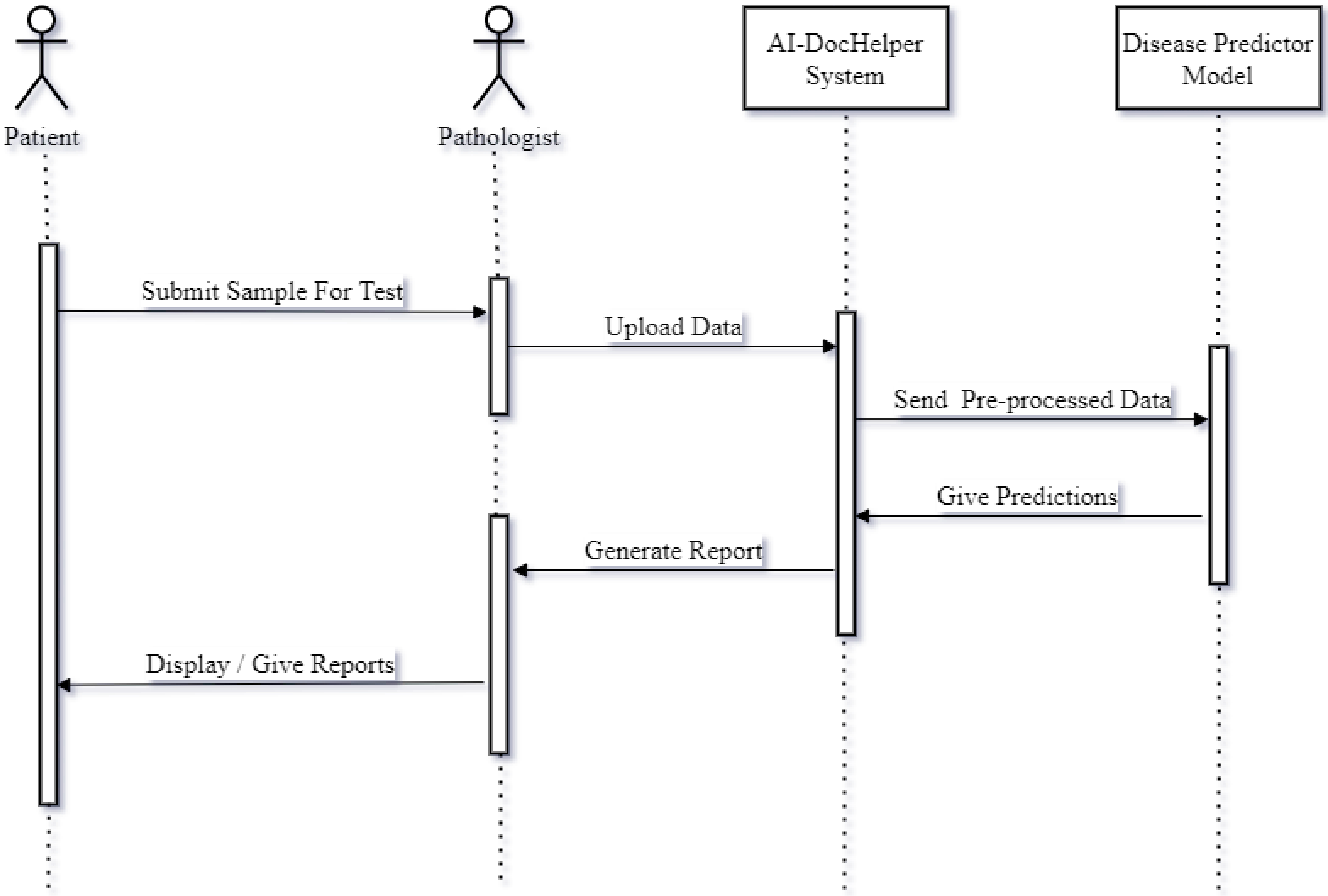


Fig 4.4: Sequence diagram for Real-Time Data

4.2.3 Data Preprocessing

4.2.3.1 Data Preprocessing for Textual Dataset

- Acquire the data and import all the crucial or necessary libraries for preprocessing. The libraries include Numpy (for matrix/scientific calculation), pandas (for data manipulation or analysis), and matplotlib (for plotting all the graphs and charts)
- Import the dataset(s) that is to be preprocessed.

- Identify the missing values and handling them. Handling it involves either dropping all the rows that are having missing values. This method will reduce the size of the dataset. The other option is to take the mean of the entire column, that contains the missing values, and filling it with the value acquired. Other methods involve, taking the median, mode, or standard deviation.
- Encode the categorical data. Categorical data refers to the information that has specific categories within the dataset.
- Splitting the dataset into training and testing set. Training set denotes the subset of a dataset that is used for training the machine learning model. Here, you are already aware of the output. A test set, on the other hand, is the subset of the dataset that is used for testing the machine learning model. The ML model uses the test set to predict outcomes.
- Feature scaling marks the end of the data preprocessing phase. It is a method to standardize the independent variables of a dataset within a specific range. In other words, feature scaling limits the range of variables so that you can compare them on common grounds.

4.2.3.2 Data Preprocessing for Image Dataset

- First, all images are going to be either resized to 256 by 256 pixels large or 100 by 100 pixels large. 256 by 256 size is considered for all the grayscale images and 100 by 100 size is considered for all the color images. This is important to do since the images in all folders are have different dimensions while the neural networks can only accept data with fixed array size.
- All the X-ray or CT scans are in grayscale format. But, the images that are available with us are in RGB format. We can train the system in RGB format, but it will increase the overhead and take an ample amount of time to be trained. If these images are trained in grayscale format, it will reduce the overhead time, give better accuracy, and will take less space in the system memory once it is stored.
- We perform normalization to reduce the effect of illumination differences. Moreover, the CNN converges faster on [0...1] data than on [0...255].
- As the data seems imbalanced, to increase the no. of training examples, we will use data augmentation. To avoid the overfitting problem, we need to expand our dataset artificially. We are making the existing dataset even larger. The idea is to alter the training data with small transformations to reproduce the variations. Approaches that alter the training data in ways that

change the array representation while keeping the label the same are known as data augmentation techniques. Some popular augmentations people use are grayscales, horizontal flips, vertical flips, random crops, color jitters, translations, rotations, and much more. By applying just a couple of these transformations to our training data, we can easily double or triple the number of training examples and create a very robust model. The main point of augmenting data — or more specifically augmenting train data is that we are going to increase the number of data used for training by creating more samples with some sort of randomness on each of them. These randomnesses might include translations, rotations, scaling, shearing, and flips. Such a technique can help our neural network classifier to reduce overfitting, or in other words, it can make the model generalize data samples better. The implementation is very easy thanks to the existence of *ImageDataGenerator* object which can be imported from *the Keras* module.

4.2.4 Comparative Study for Textual Dataset

For selecting a model for the system, we have conducted a comparative study. This comparative study was done to identify which machine learning or deep learning model is to be selected for the classification of a particular disease if the data is available in textual format.

For this purpose, we have selected a dataset for Diabetes classification. This diabetes classification data, also known as Pima Indian Diabetes Database or Dataset, was acquired from Kaggle, which was published by UCI Machine Learning repository.

The dataset was originally from the National Institute of Diabetes and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on some of the diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Dataset Link : https://www.kaggle.com/uciml/pima-indians-diabetes-database

This dataset has the following fileds :

- Pregnancies : Number of times pregnant
- Glucose : Plasma glucose concentration a 2 hours in an oral glucose tolerance test

- BloodPressure : Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skinfold thickness (mm)
- Insulin : 2-Hour serum insulin (mu U/ml)
- BMI : Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction : Diabetes pedigree function (It gives the history of diabetes in the family)
- Age : Age (years)
- Outcome : Class variable (0 or 1)

This dataset was pre-processed and fed to many machine learning and deep learning models such as Logistic Regression, K-Nearest Neighbour Classifier, Support Vector Classifier, Naïve Bayes, Decision Tree Classifier, Random Forest Classifier, and Artificial Neural Network, to decide which machine or deep learning model gives the best accuracy and has the least loss. The confusion matrix along with the accuracy of each model or algorithm is given below in fig 4.5.

Table 4.1, the table below, shows the accuracy of different machine learning and deep learning models. Here, we can see that Logistic regression, Support Vector Classifier, and Random Forest Classifier have given the same accuracy. So, we have chosen Random Forest Classifier for all the textual datasets.

MODEL	**ACCURACY**
Logistic Regression	82.46%
K-Nearest Neighbour Classifier	79.87%
Support Vector Classifier	82.46%
Naïve Bayes	79.22%
Decision Tree Classifier	70.77%
Random Forest Classifier	82.46%
Artificial Neural Network	70.77%

Table 4.1 : Comparision of Accuracy for Textual Dataset

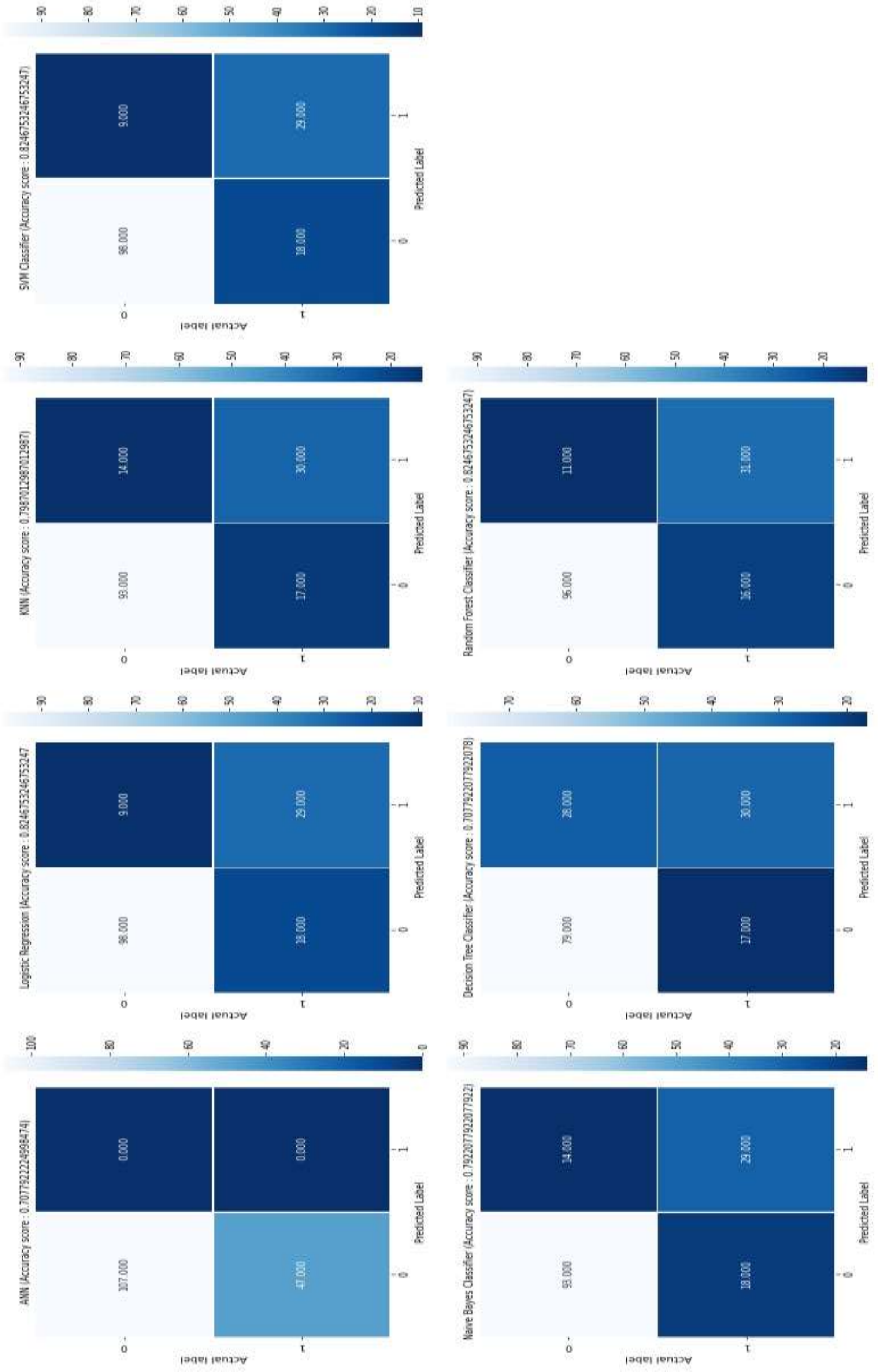


Fig 4.5: Confusion Matrix for Diabetes Detector with Accuracy

4.2.5 Study for Image Dataset

For all the image data or datasets available for disease detection, our system will make use of Convolutional Neural Networks, which are a part of deep learning. All the image data will be preprocessed to extract all the relevant features for the detection of a particular disease. All the irrelevant features or data will be removed from the images. The irrelevant data will be removed because it will affect the accuracy of the model and will make it difficult for the presence of that particular disease.

4.2.6 Algorithm

4.2.6.1 Random Forest Classifier (Textual Dataset)

Random Forest Classifier is going to be used for classifying all the textual data. Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science-speak, the reason that the random forest model works so well is ***A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.***

The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this wonderful effect is that the trees protect each other from their errors (as long as they don't constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees can move in the correct direction. So the prerequisites for the random forest to perform well are:

1. There needs to be some actual signal in our features so that models built using those features do better than random guessing.

2. The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

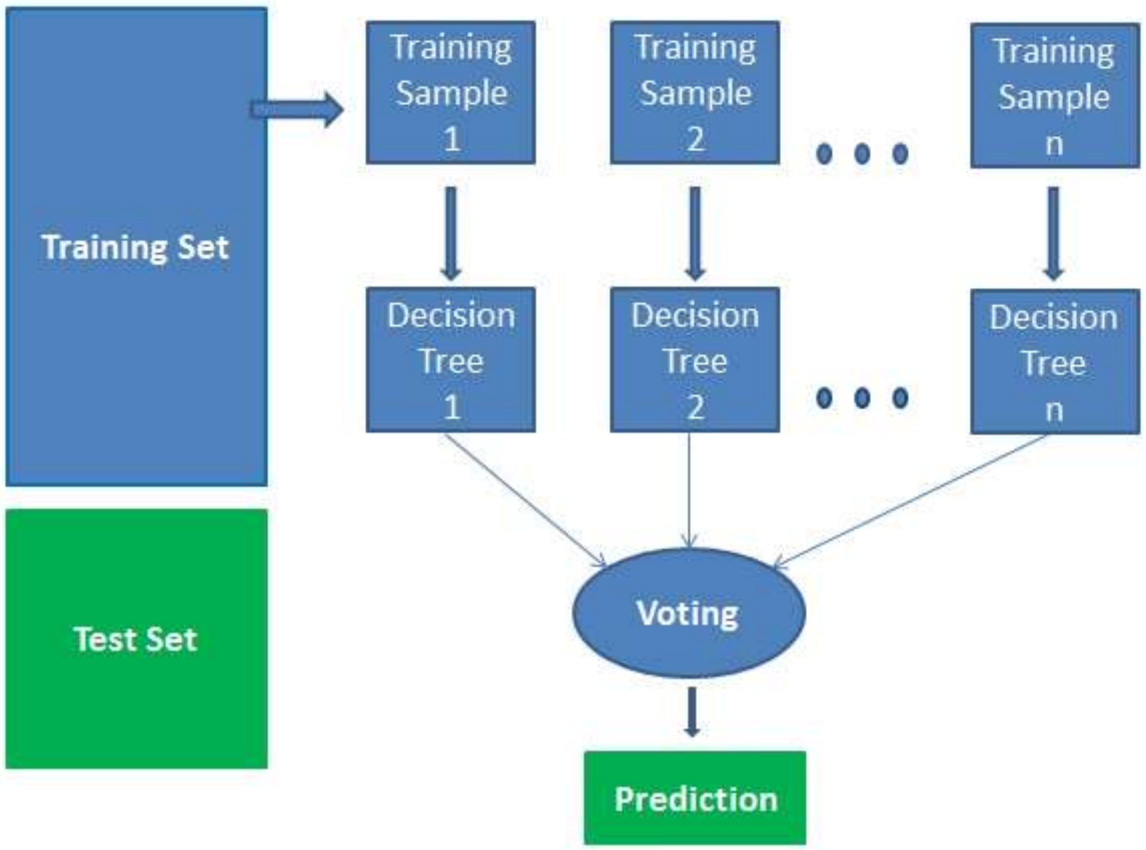


Fig 4.6: Random Forest Classifier

It works in four steps:

- Select random samples from a given dataset.
- Construct a decision tree for each sample and get a prediction result from each decision tree.
- Perform a vote for each predicted result.
- Select the prediction result with the most votes as the final prediction.

4.2.6.2 Convolutional Neural Network (CNN) (Image Dataset)

Convolutional Neural Network (CNN) is going to be applied to all the image data. The image data can either be in Grayscale or RGB format. For Grayscale images, a Custom CNN will be built to extract all the features, whereas a CNN model with Transfer Learning will be used to extract all the data from RGB images.

Convolutional Neural Network (CNN) is a Deep Learning algorithm that can take in an input image, assign importance (learnable weights and biases) to various aspects in the image, and be able to differentiate one from the other. The reason for using CNN is because of its ability to successfully capture the Spatial and Temporal dependencies in an image through the application of relevant filters. The architecture performs a better fitting to the image dataset due to the reduction in the number of parameters involved and the reusability of weights. In other words, the network can be trained to understand the sophistication of the image better. The model training starts by passing the image through a series of convolutional, nonlinear, pooling layers and fully connected layers and then generates the output.

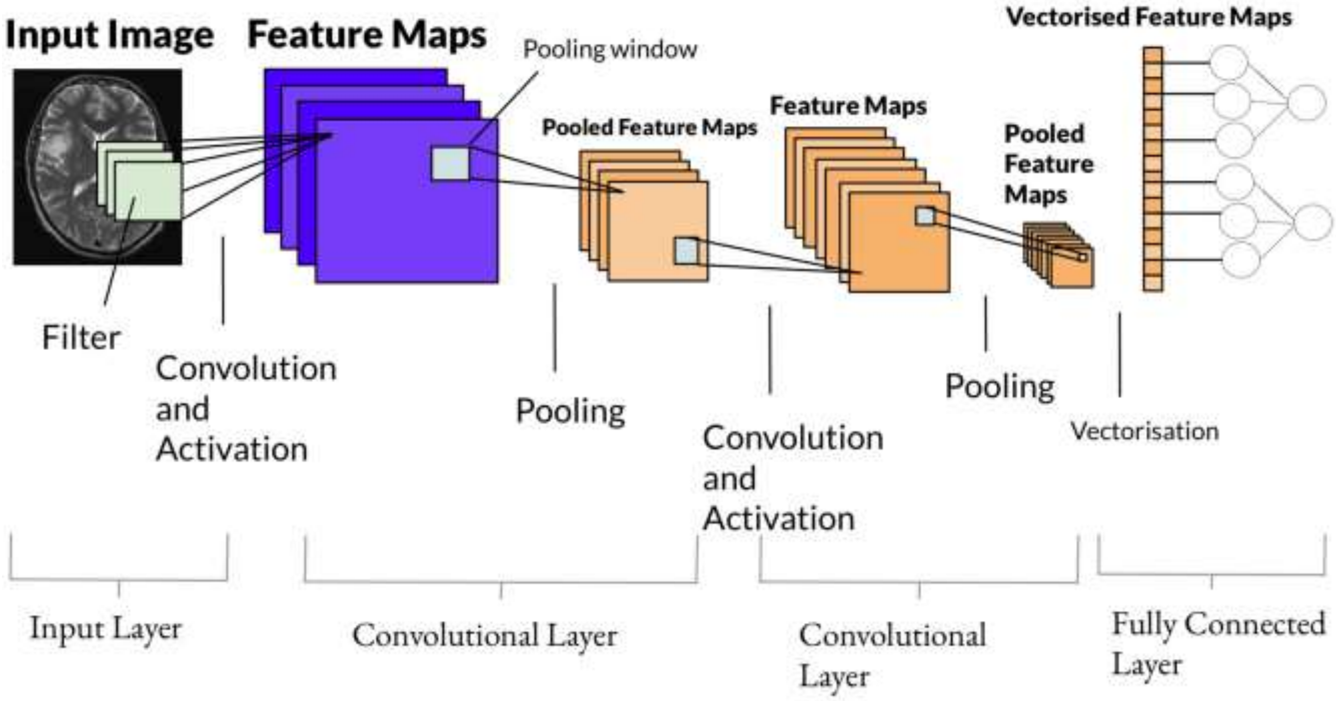


Fig 4.7: Convolutional Neural Network

The Convolution layer is always the first. The image is entered into it. Then the filter produces convolution, i.e. moves along the input image. The network will consist of several convolutional networks mixed with nonlinear and pooling layers. When the image passes through one convolution layer, the output of the first layer becomes the input for the second layer. And this happens with every further convolutional layer.

The nonlinear layer is added after each convolution operation. It has an activation function, which brings nonlinear property. Without this property, a network would not be sufficiently intense and will not be able to model the response variable (as a class label).

The pooling layer follows the nonlinear layer. It works with the width and height of the image and performs a downsampling operation on them. As a result, the image volume is reduced. After completion of a series of convolutional, nonlinear, and pooling layers, it is necessary to attach a fully connected layer. This layer takes the output information from convolutional networks. Attaching a fully connected layer to the end of the network results in a 2-dimensional vector denoting if the disease is present or not.

4.3 GUI Design

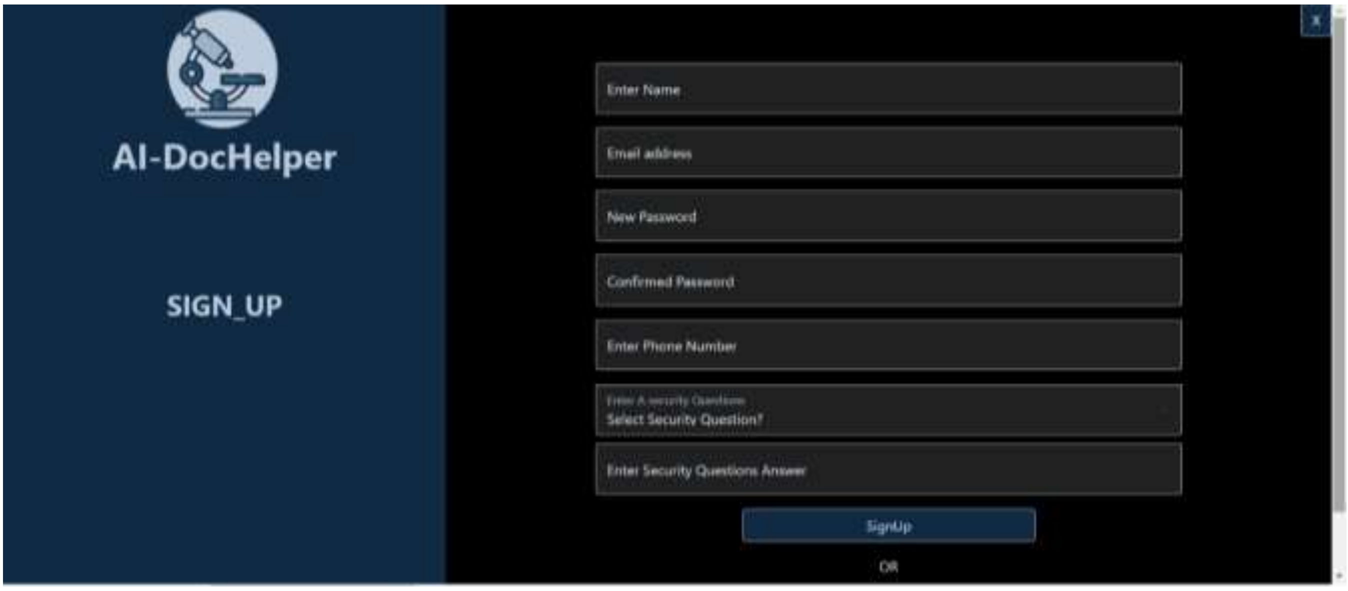


Fig 4.8: Sign Up Page

In Fig. 4.8, shown above, a user can register into the system by providing all the necessary required details.

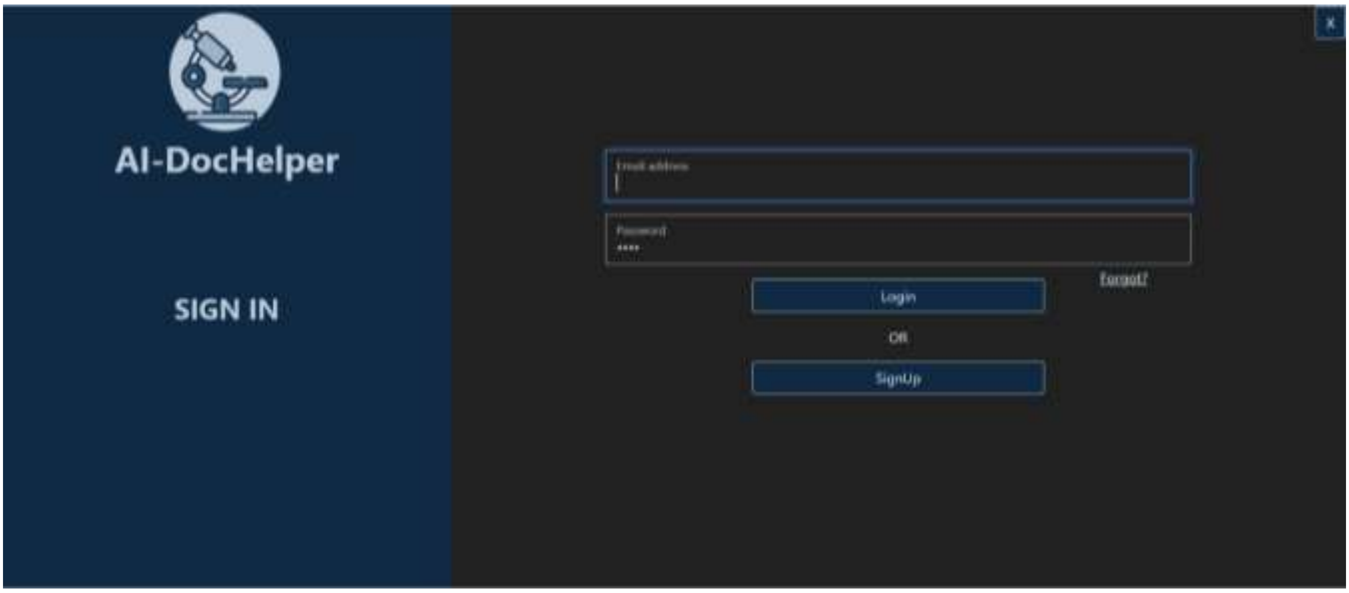


Fig 4.9: Sign In Page

In Fig. 4.9, shown above, a user can log in to the system by the system by providing all the necessary details. If the user forgets the password, the user can click the Forgot button, and reset its password. Once the user clicks the Forgot? Button, the user will be redirected to the Forgot Password page, as shown in Fig. 4.10.

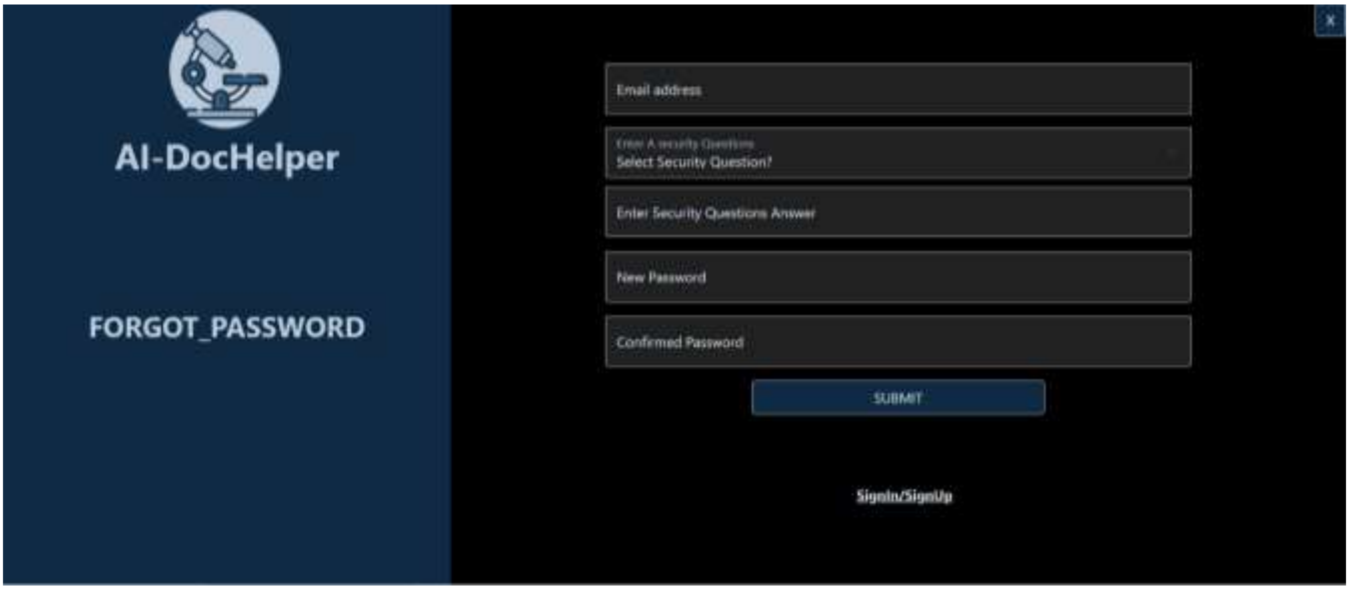
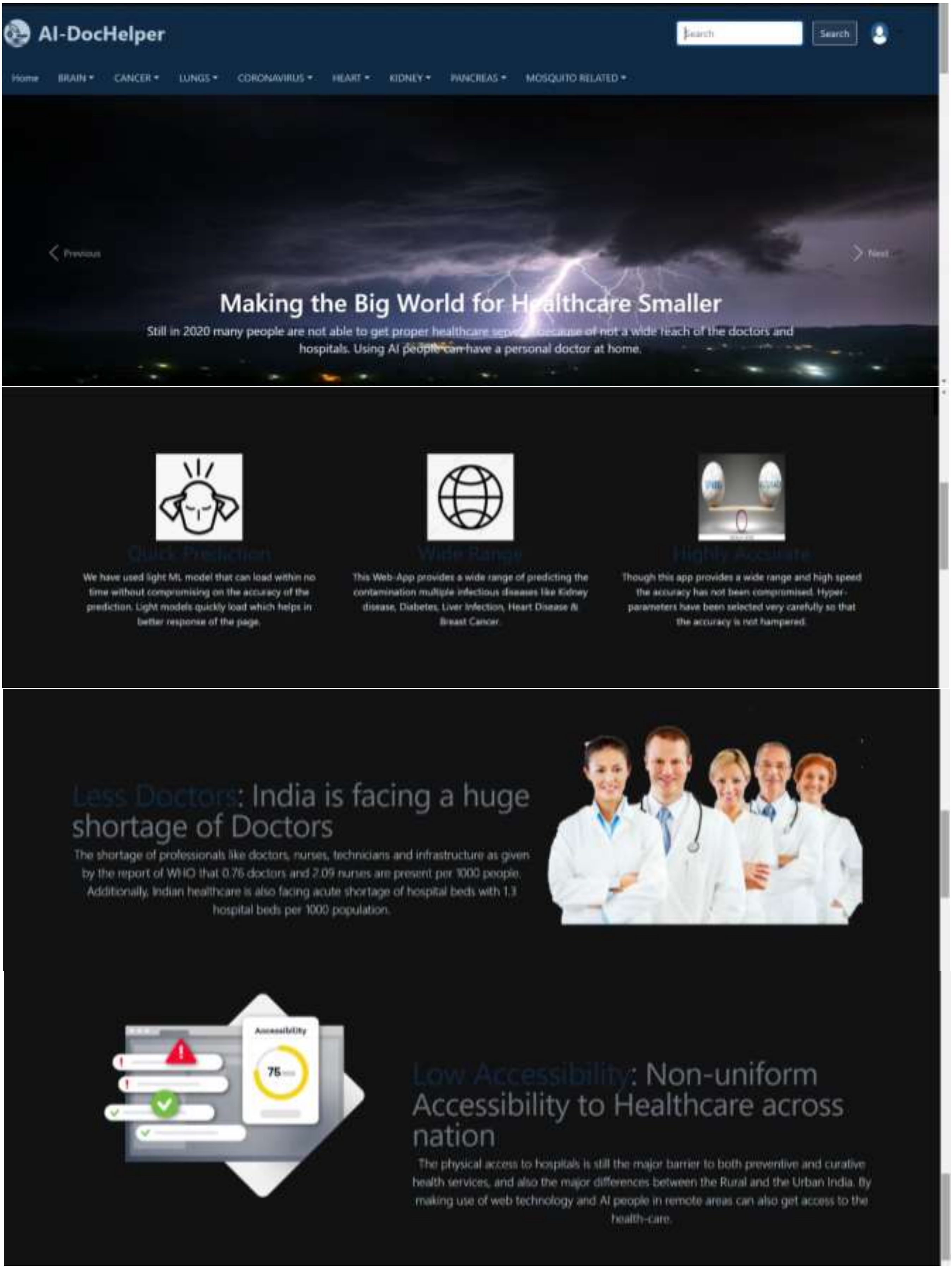


Fig 4.10: Forgot Password Page



Making the Big World for Healthcare Smaller

Still in 2020 many people are not able to get proper healthcare services because of not a wide reach of the doctors and hospitals. Using AI people can have a personal doctor at home.

Quick Prediction

We have used light ML model that can load within no time without compromising on the accuracy of the prediction. Light models quickly load which helps in better response of the page.

Wide Range

This Web-App provides a wide range of predicting the contamination multiple infectious diseases like Kidney disease, Diabetes, Liver Infection, Heart Disease & Breast Cancer.

Highly Accurate

Though this app provides a wide range and high speed the accuracy has not been compromised. Hyper-parameters have been selected very carefully so that the accuracy is not hampered.

Less Doctors: India is facing a huge shortage of Doctors

The shortage of professionals like doctors, nurses, technicians and infrastructure as given by the report of WHO that 0.76 doctors and 2.09 nurses are present per 1000 people. Additionally, Indian healthcare is also facing acute shortage of hospital beds with 1.3 hospital beds per 1000 population.

Low Accessibility: Non-uniform Accessibility to Healthcare across nation

The physical access to hospitals is still the major barrier to both preventive and curative health services, and also the major differences between the Rural and the Urban India. By making use of web technology and AI people in remote areas can also get access to the health-care.

Fig 4.11: Home Page

Fig. 4.11, shown above displays the Home page. The user can navigate throughout the system using the page. The user can navigate to any part of the system from here, provided the user has logged in. If the user isn't logged in, its functionality cannot be accessed. The user is also provided with a search bar, to search for any disease and it will be redirected to that disease detection page.

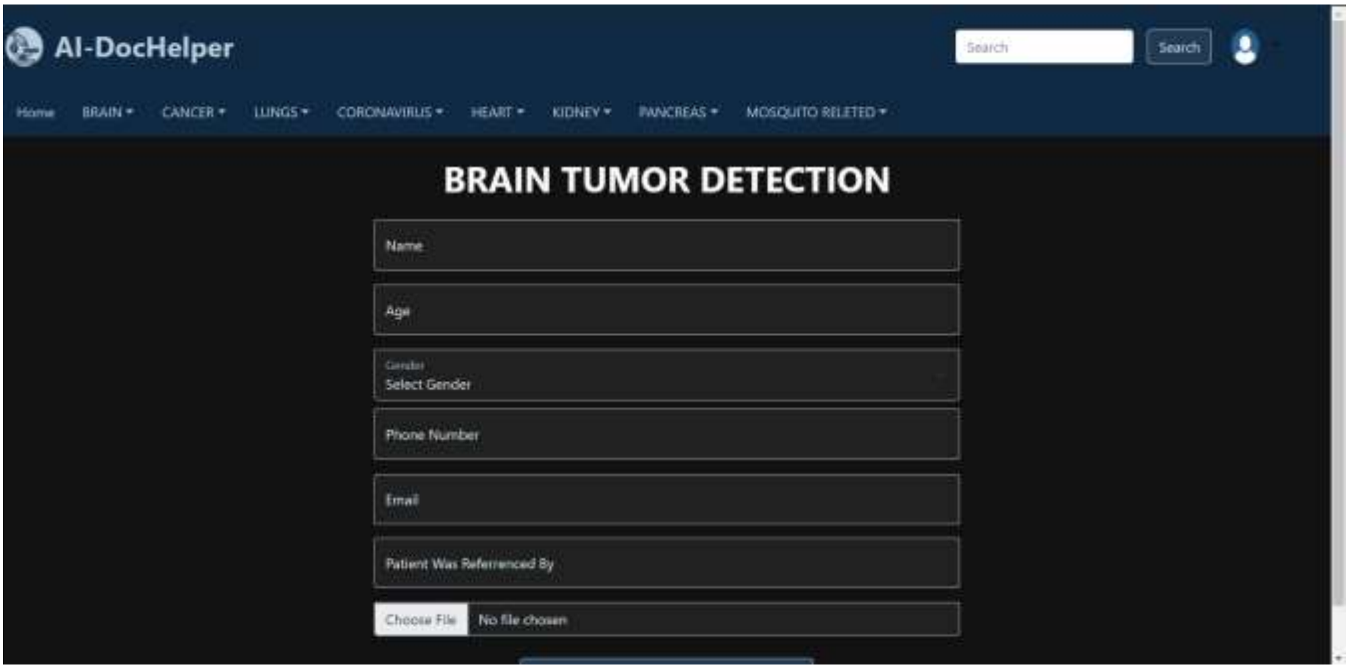
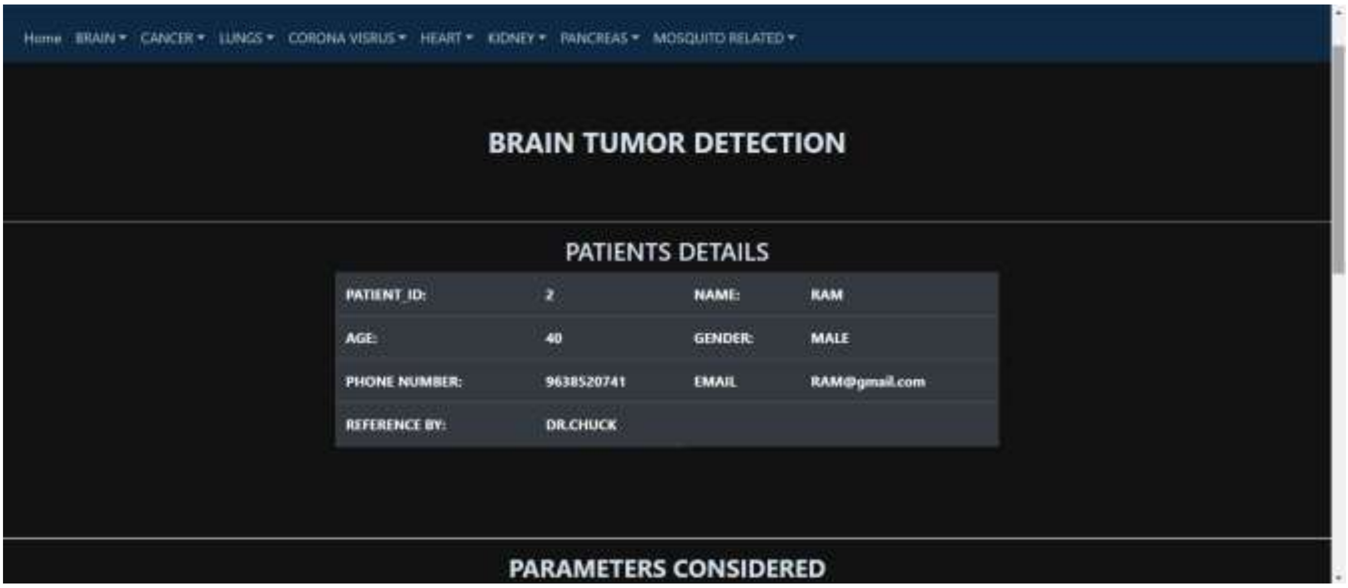


Fig 4.12: Input Page for Image Dataset

Fig. 4.12, shown above, will be the UI for all the diseases that can be detected using/ with the help of images. Once all the necessary data is entered correctly, the user will be redirected to the result page that shows the result/report for the disease detected, shown in Fig 4.13.



PATIENTS DETAILS

PATIENT_ID:	2	NAME:	RAM
AGE:	40	GENDER:	MALE
PHONE NUMBER:	9638520741	EMAIL	RAM@gmail.com
REFERENCE BY:	DR.CHUCK		

PARAMETERS CONSIDERED

Fig 4.13: Result Page for Image Dataset

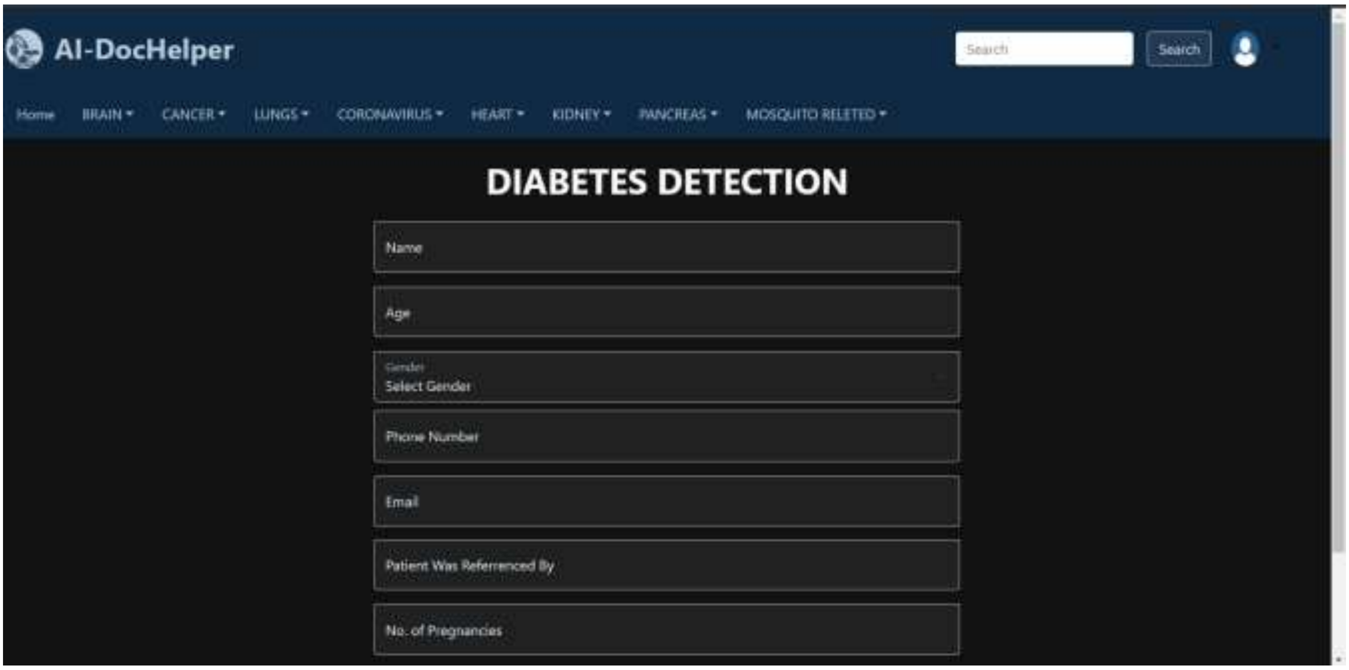


Fig 4.14: Input Page for Textual Dataset

Fig. 4.14, shown above, will be the UI for all the diseases that can be detected using/with the help of text as the input. Once all the necessary data is entered correctly, the user will be redirected to the result page that shows the result/report for the disease detected.

Chapter 5

Implementation

5.1 Plan of Implementation

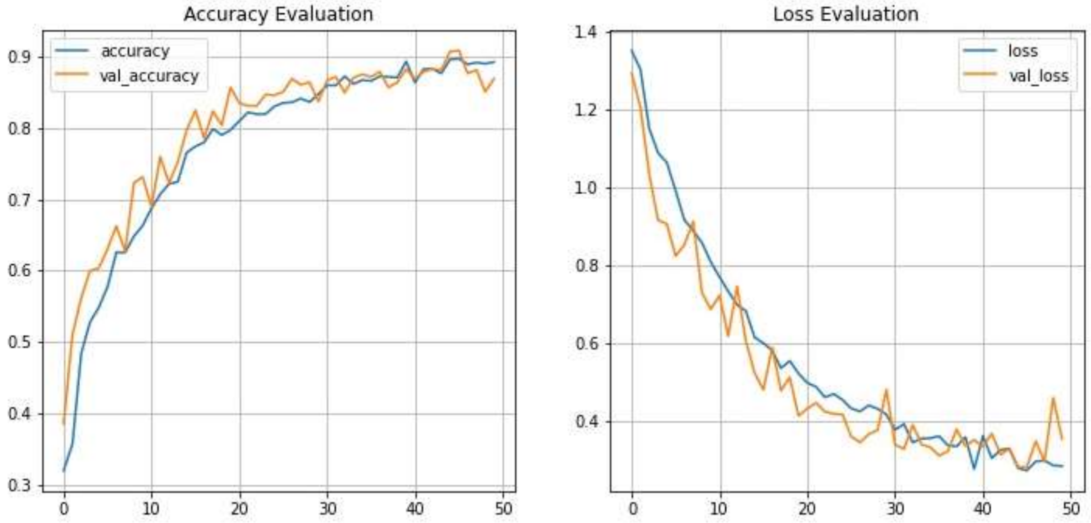
Date	Task	Description
01/11/2020 – 01/12/2020	Dataset Research	Research for the textual and image dataset for diseases. All the data cleaning and preprocessing work are also done.
01/11/2020 – 15/112020	Literature Survey	A survey for the existing systems was done and the limitations were noted.
02/12/2020 to 31/12/2020	Algorithm Testing	All the algorithms will be tested on the image and text dataset, and the model with the best accuracy and least loss will be selected.
01/01/21 to 15/02/21	Implementation of Model	Once the algorithm is selected, all the datasets will be trained on the model. The model will also be saved for future use.
16/02/21 to 30/02/21	Implementation of GUI	A user-friendly GUI will be implemented on the finalized design.
01/03/21 to 30/03/21	First Testing	The front-end and the back-end will be connected and the system will be tested. All the features of the system will be tested.
01/04/21 to 15/04/21	Deployment Phase	Bugs will be fixed and the model will be made ready for deployment. The GUI and database will be deployed on the server.
16/04/21 to 30/04/21	Final Testing	Testing will be done on the UI and backend. Bugs will be fixed. Once all the bugs are fixed, testing will be done. This ends the final testing phase.

Table 5.1: Plan of Implementation

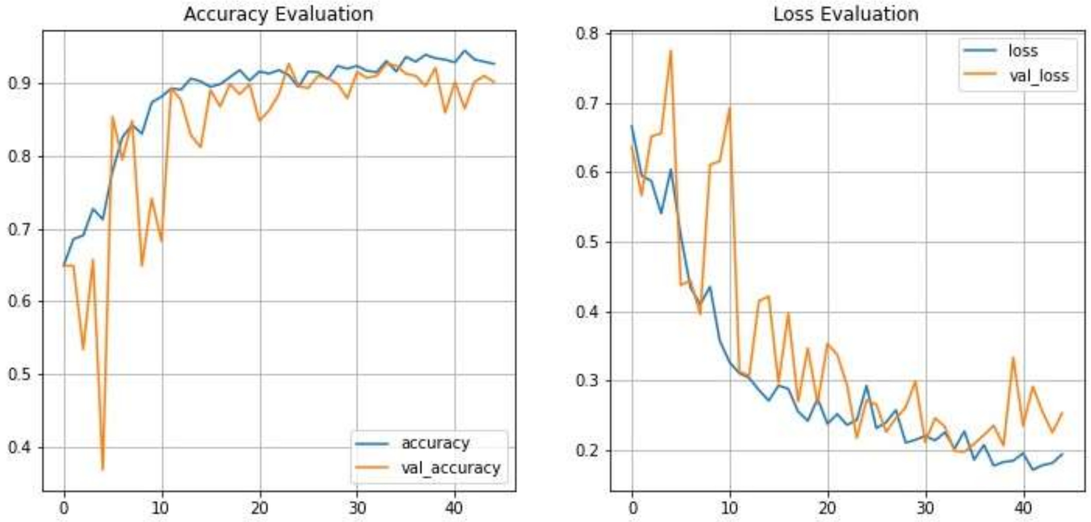
5.2 Results And Evaluation

Disease Name	Training Accuracy	Training Loss	Testing Accuracy	Testing Loss
Breast Cancer	87.36%	31.24%	85.49%	38.67%
Pneumonia	95.86%	10.92%	98.25%	14.88%
Glioma Tumor	92.20%	19.64%	92.41%	19.73%
Meningioma Tumor	93.19%	18.18%	90.52%	21.80%
Diabetes	0.8240	0.2310	0.8052	0.2513
Pituitary Tumor	99.06%	3.33%	98.85%	3.03%
Melanoma (Skin) Cancer	85.12%	31.37%	83.87%	33.26%
Acute Lymphoblastic Leukemia (ALL)	84.14%	37.54%	70.43%	58.85%
Heart Disease	0.9137	0.2513	0.8897	0.2883
Chronic Kidney	0.8963	0.1953	0.8663	0.2153
Malaria	89.69%	26.80%	90.49%	24.19%
Brain Tumor	89.21%	28.49%	66.67%	33.72%
Lung Cnacer	94.41%	13.80%	95.26%	11.60%
Invasive Ductal Carcinoma (IDC)	80.57%	43.16%	78.70%	47.90%
Tuberculosis	96.69%	9.75%	95.64%	11.64%
COVID-19	86.56%	33.12%	86.90%	32.88%

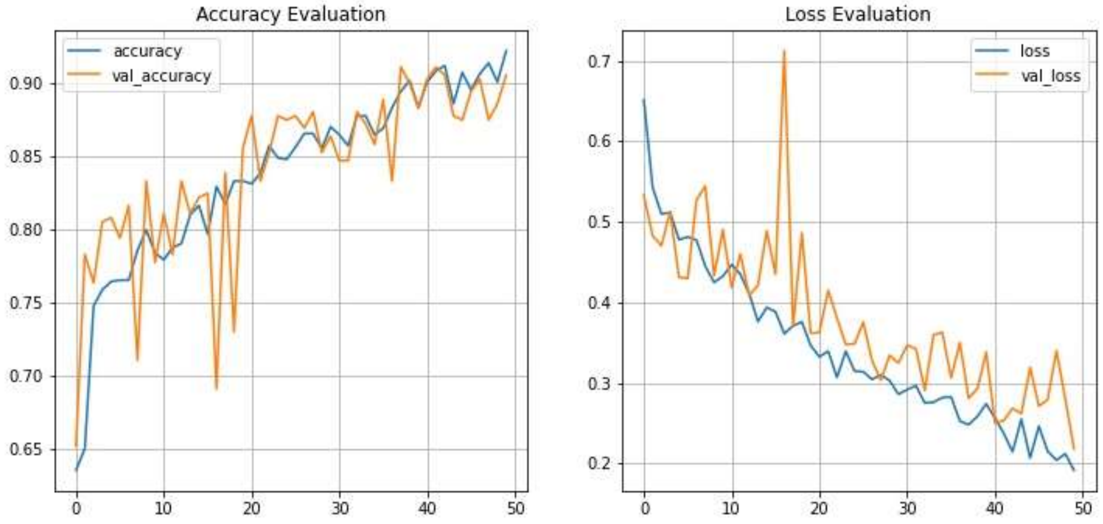
Table 5.2: Accuracy & Loss of All Models



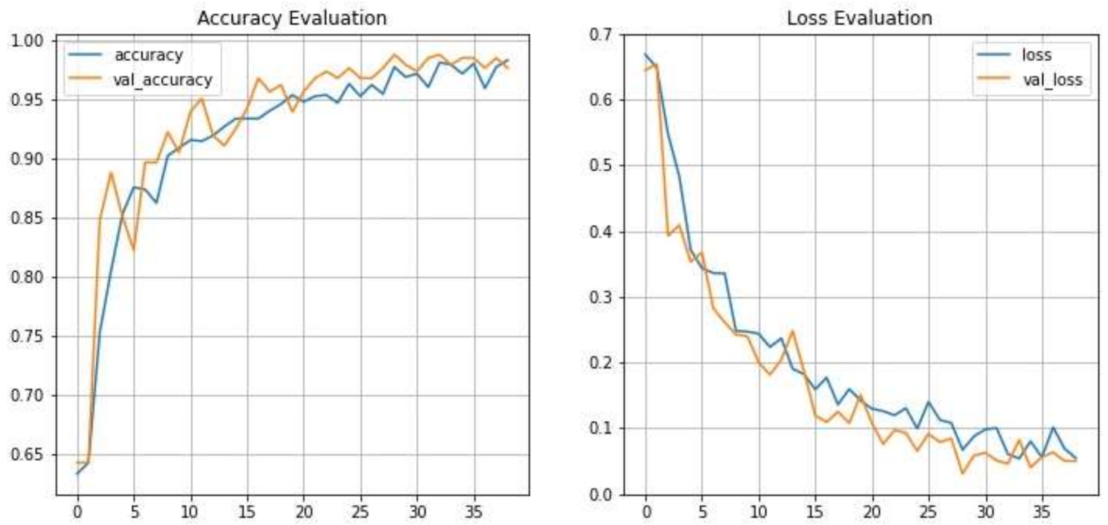
Brain Tumor Classification



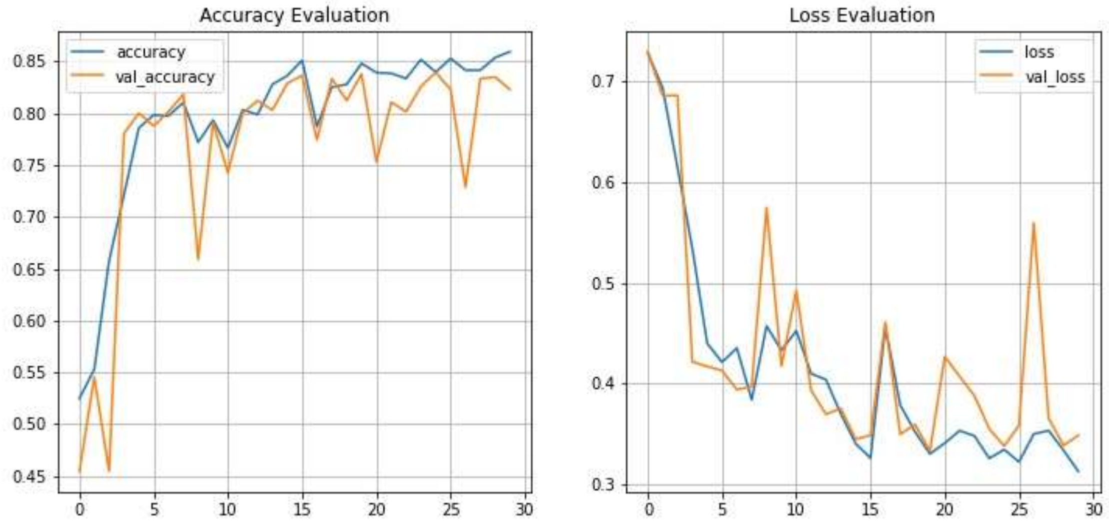
Giloma Tumor Classification



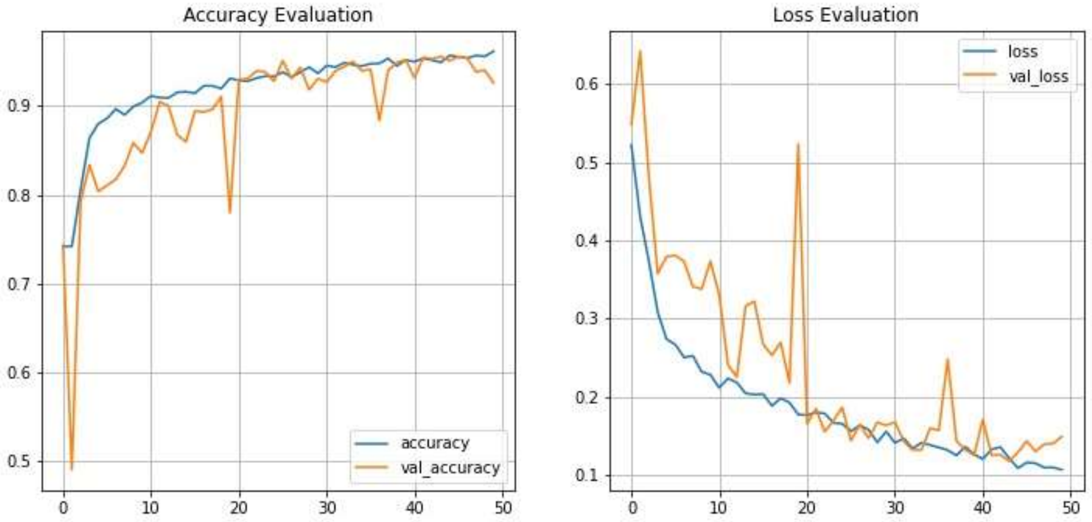
Meningioma tumor Classification



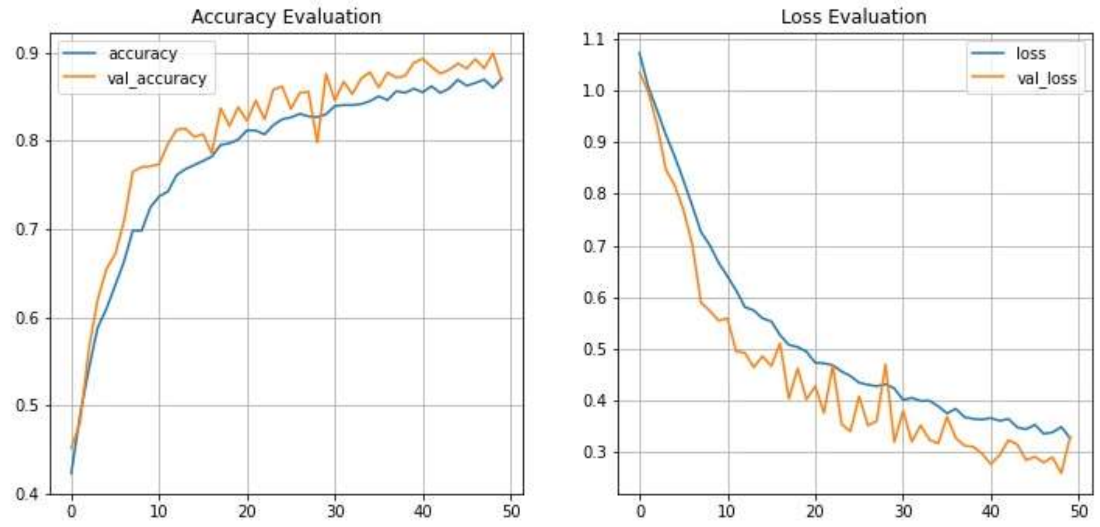
Pitutitary Tumor Classification



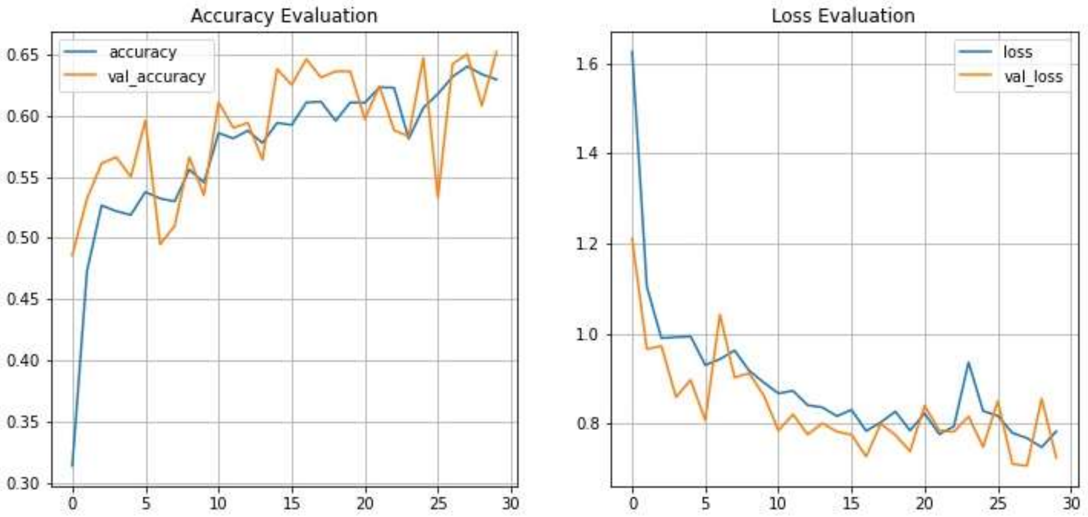
Melanoma (Skin) Classification



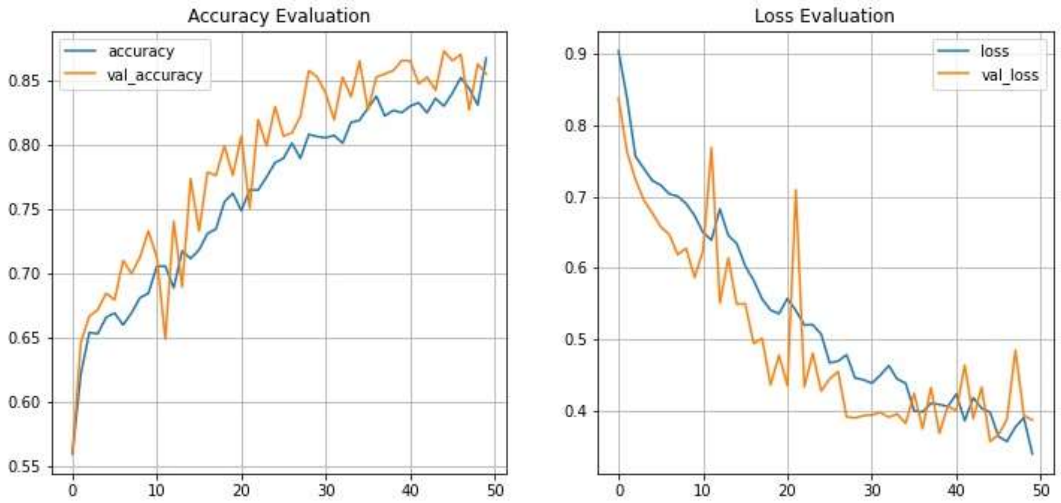
Pneumonia Detection



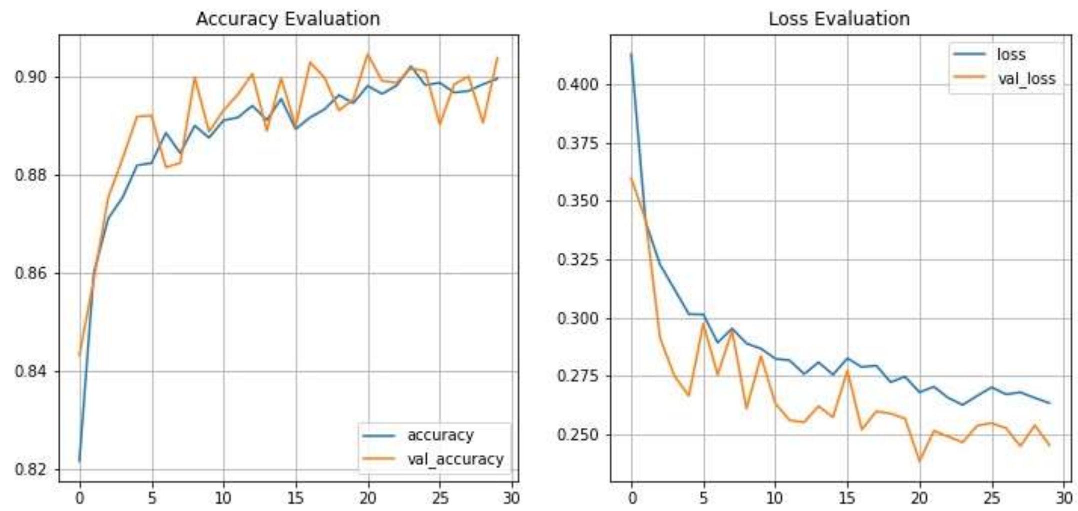
COVID-19 Detection



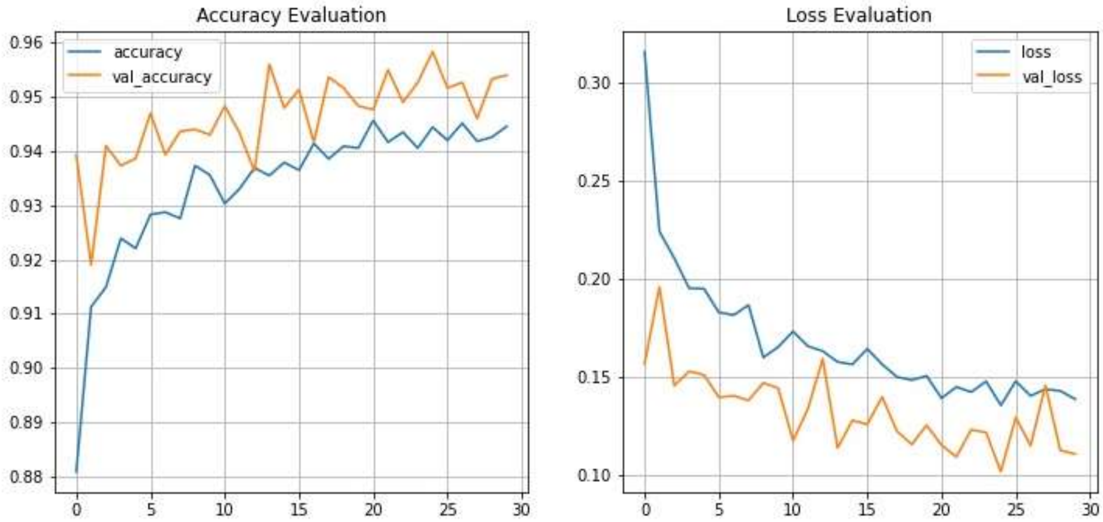
Colon Cancer Detection



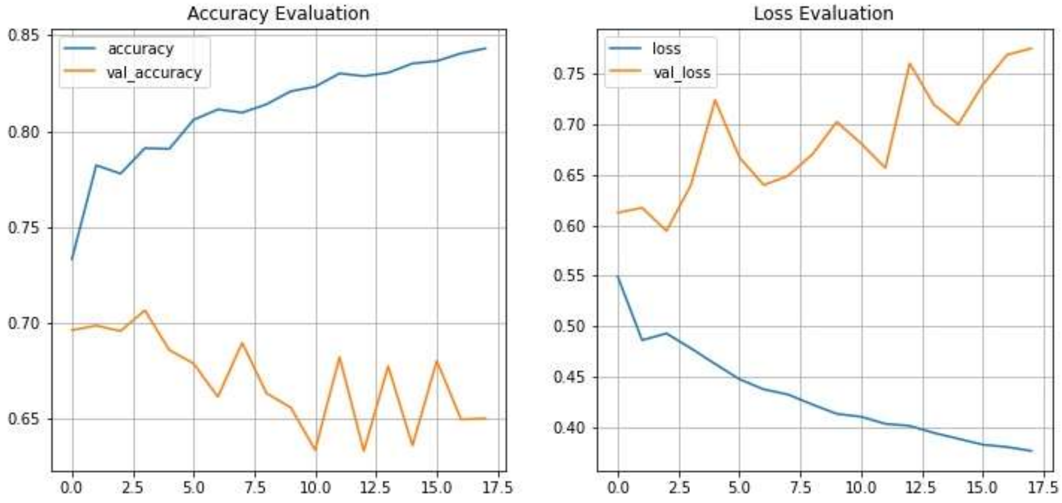
Breast Cancer Detection



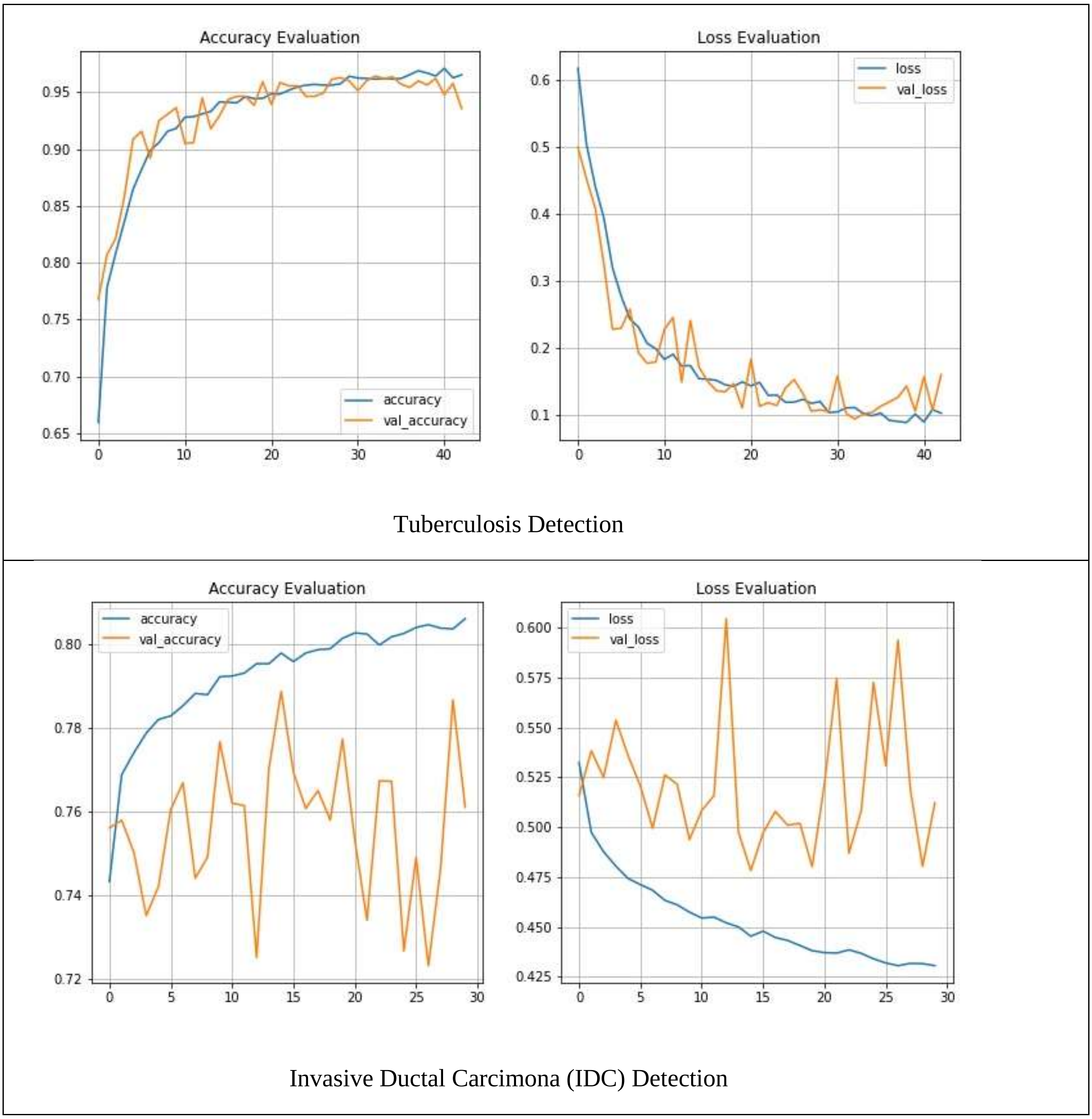
Malaria Detection



Lung Cancer Detection



Acute Lymphobiastic Leukemia (ALL) Detection



Tuberculosis Detection

Invasive Ductal Carcimona (IDC) Detection

Fig 5.1: Accuracy Graph and Loss Graph of Models

Chapter 6

Conclusion

The research was done, and no system was found available. The available systems had restrictions that they were able to find a particular type of disease. Available systems are restricted to find only one or two diseases.

AI-DocHelper is better than many systems available in the market as it can detect many diseases. Once the system is available in the market, it will help to save many lives with precision. Human errors will be reduced to a great extent. A doctor can plan the course of treatment for a patient very quickly, as the system gives the results immediately, as compared to modern systems available.

Once more datasets are available, they can be trained, and new diseases can be added to the system. The detection models can be trained using more data, to increase their accuracy and the loss. The more the model has trained the more its features for detection of disease will get fine-tuned.

References

1) Bagga, P.; Hans, R. Applications of Mobile Agents in Healthcare Domain: A Literature Survey. Int. J. Grid Distrib. Comput. 2015, 8. [CrossRef]

2) Grimson, J.; Stephens, G.; Jung, B.; Grimson, W.; Berry, D.; Pardon, S. Sharing health-care records over the Internet. IEEE Internet Comput. 2001, 5, 49–58. [CrossRef]

3) Daniels, M.; Schroeder, S.A. Variation among physicians in use of laboratory tests II. Relation to clinical productivity and outcomes of care. Med. Care 1977, 482–487. [CrossRef] [PubMed]

4) Wennberg, J.E. Dealing with medical practice variations: A proposal for action. Health Aff. 1984, 3, 6–32. [CrossRef]

5) Smellie, W.S.A.; Galloway, M.J.; Chinn, D.; Gedling, P. Is clinical practice variability the major reason for differences in pathology requesting patterns in general practice? J. Clin. Pathol. 2002, 55, 312–314. [CrossRef]

6) Stuart, P.J.; Crooks, S.; Porton, M. An interventional program for diagnostic testing in the emergency department. Med J. Aust. 2002, 177, 131–134. [CrossRef]

7) Pölsterl, S.; Conjeti, S.; Navab, N.; Katouzian, A. Survival analysis for high-dimensional, heterogeneous medical data: Exploring feature extraction as an alternative to feature selection. Artif. Intell. Med. 2016, 72, 1–11. [CrossRef]

8) Dick, R.S.; Steen, E.B.; Detmer, D.E. The Computer-Based Patient Record: An Essential Technology for Health Care; National Academies Press: Washington, DC, USA, 1997.

9) Zhuang, Z.Y.; Churilov, L.; Burstein, F.; Sikaris, K. Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners. Eur. J. Oper. Res. 2009, 195, 662–675. [CrossRef]

10) Huang, M.J.; Chen, M.Y.; Lee, S.C. Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. Expert Syst. Appl. 2007, 32, 856–867. [CrossRef]

11) R. Sangeetha and K. S. Murthy, "A novel approach for detection of breast cancer at an early stage using digital image processing techniques," *2017 International Conference on Inventive Systems and Control (ICISC)*, Coimbatore, 2017, pp. 1-4, doi: 10.1109/ICISC.2017.8068625.

12) Y. Lu, J. Li, Y. Su and A. Liu, "A Review of Breast Cancer Detection in Medical Images," 2018 IEEE Visual Communications and Image Processing (VCIP), Taichung, Taiwan, 2018, pp. 1-4, doi: 10.1109/VCIP.2018.8698732.

13) S. Nayak, S. Kumar and M. Jangid, "Malaria Detection Using Multiple Deep Learning Approaches," 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur, India, 2019, pp. 292-297, doi: 10.1109/ICCT46177.2019.8969046.

14) X. Zeng, H. Chen, Y. Luo and W. Ye, "Automated Diabetic Retinopathy Detection Based on Binocular Siamese-Like Convolutional Neural Network," in IEEE Access, vol. 7, pp. 30744-30753, 2019, doi: 10.1109/ACCESS.2019.2903171.

15) T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins and R. H. Ghomi, "Parkinson's Disease Diagnosis Using Machine Learning and Voice," 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), Philadelphia, PA, 2018, pp. 1-7, doi: 10.1109/SPMB.2018.8615607.

16) Y. Liu et al., "Detecting Diseases by Human-Physiological-Parameter-Based Deep Learning," in IEEE Access, vol. 7, pp. 22002-22010, 2019, doi: 10.1109/ACCESS.2019.2893877.

17) A. Shrivastava, I. Jaggi, S. Gupta and D. Gupta, "Handwritten Digit Recognition Using Machine Learning: A Review," 2019 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC), Greater Noida, India, 2019, pp. 322-326, doi: 10.1109/PEEIC47157.2019.8976601.

Acknowledgment

Every project big or small is successful largely due to the effort of wonderful people who have given their valuable advice or lent a helping hand. We sincerely appreciate the inspiration, support, and guidance for all those people that have been instrumental in making this project a success.

We take this opportunity to express our profound gratitude and deep regard to our project guide “**Mr. Aejaz Khan**” for his exemplary guidance, monitoring, and constant encouragement. The blessing, help, and guidance have given by him time to time shall carry us in the long way of our journey which we are about to embark on.

We are also obliged to all the faculty members of the computer department, for the valuable information provided by them. We are grateful for their cooperation.

We also acknowledge the deep sense of relevance, our gratitude towards our parents and family, who has always supported us morally as well as economically. Without them, this project would not be possible.

Last but not the least, we are indebted to our college “**Thadomal Shahani Engineering College**” for giving us the platform to express and exhibit our talent.