**PROJECT REPORT**

##### **on**

**Histopathologic Breast Cancer Detection**

###### **for**

**Digipodium**

**towards partial fulfillment of the requirement**

**for the award of degree of**

###### **Bachelor of Computer Applications**

**from**

# **Babu Banarasi Das University**

**Lucknow**

**Developed and Submitted by Under Guidance of**

Falguni Jiya

Ayush Tiwari Zaid Kamil

**Academic Session 2018 - 19**

### **School of Computer Applications**

|  |
| --- |
| **I Floor, H-Block, BBDU, BBD City, Faizabad Road, Lucknow (U. P.) INDIA 226028**  **PHONE: HEAD:** 0522-3911127, 3911321 **Dept. Adm. & Exam Cell:** 0522-3911326 **Dept. T&P Cell:** 0522-3911128; **E-Mail:** head.sca@gmail.com |
| **[www.bbdu.ac.in](http://www.bbdu.ac.in/)** |

# **Babu Banarasi Das University**

**Lucknow**

**CERTIFICATE**

###### **This is to certify that Project Report entitled**

**Histopathologic Breast Cancer Detection**

**being submitted by**

**Falguni Jiya And Ayush tiwari**

**towards the partial fulfillment of the requirement**

**for the award of the degree of**

###### **Bachelor of Computer Applications**

**to**

# **Babu Banarasi Das University**

**Lucknow**

**in the Academic Year 2018-19**

**is a record of the student’s own work carried out at**

### **Digipodium**

**and to the best of our knowledge the work reported herein does not form a part of any other thesis or work on the basis of which degree or award was conferred on an earlier occasion to this or any other candidate.**

#### 

#### **Prabhash Ch. Pathak**

**HEAD (School of Computer Applications)**

**ACKNOWLEDGEMENT**

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to DIGIPODIUM INSTITUTE for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards my friends & member of DIGIPODIUM for their kind co-operation and encouragement which help me in completion of this project.

I would like to express my special gratitude and thanks to group members for giving me such attention and time.

My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities.

**DECLARATION**

I **FALGUNI JIYA, AYUSH TIWARI** hereby declare that this project report entitled **HISTOPATHOLOGIC BREAST CANCER DETECTION**, submitted by us, under the guidance of **Mr.** **ZAID KAMIL SIR** of **DIGIPODIUM INSTITUTE, LUCKNOW** is our own and has not been submitted to any other University or Institute or published earlier.

**Signature of Student** **:**

**FALGUNI JIYA**

**AYUSH TIWARI**

**BCA VIth Semester**

**Date: 26- April- 2019**

**ABSTRACT**

Cancer is the name given to a collection of related diseases. In breast cancer, some of the breast’s cells begin to divide without stopping and spread into surrounding tissues.Breast cancer can start in women of age group 45-60. Normally, breast cells grow and divide to form new cells as the body needs them. When cells grow old or become damaged, they die, and new cells take their place.When cancer develops, however, this orderly process breaks down. As breast cells become more and more abnormal, old or damaged cells survive when they should die, and new cells form when they are not needed. These extra cells can divide without stopping and may form growths called tumors. In this project we are focusing on detecting breast cancer at initial stage through data mining. Here we are maintaining a very large dataset and training it so that it can become more and more accurate with time.Histopathologic Breast Cancer Detection is focused on utilizing data mining methods to analyze and process the stated oncogenic data for the detection of Onco patterns Onco-protein patterns Onco protein mutations from the biological data and detection of cancer cause, symptom, patterns from the clinical data of patient records, laboratory investigations and mammography report.

**TABLE OF CONTENTS**

1. **Introduction of the Project 8**
   1. Introduction 9
2. **Need of Identification 10**

2.1 Objective 11

* 1. Scope

1. **Problem Statement 12-13**
2. **System Analysis 14**

4.1 System Design 15

4.2 Modules 16

4.3 Module Description 17-18

1. **Software Requirement Specification(SRS) 19**
2. **Gantt Chart 20**
3. **Functional Requirement 21**
4. **Use Case Diagram 22**
5. **Software Requirement 23**
6. **Hardware Requirement**
7. **High Level Design 24**

11.1 DFD 25

11.2 0 Level 26

11.3 1 Level 27

11.4 ERD 28-29

1. **Screen Shots 30-36**
2. **Challenges 37**
3. **Limitations 38**
4. **Future Scope 39**
5. **Conclusion 40**
6. **Bibliography 41**

**INTRODUCTION OF THE PROJECT**

**Introduction:**

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today’s society.

The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning is widely recognized as the methodology of choice in BC pattern classification and forecast modelling.

Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions.

Identify metastatic tissue in histopathologic scans of lymph node sections.

**NEED**

**OF IDENTIFICATION**

**Objective:**

This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this i have used machine learning classification methods to fit a function that can predict the discrete class of new input.

1. Histopathologic Cancer Detection is fully automated system which requires basic user inputs.
2. Input of least data makes it fast to interpret the results.
3. Thus it is cost effective & easier for people to approach.
4. Also it generates highly accurate results.

**Scope:**

This project is basically aimed at.

* **Detection of Breast Cancer at early stage -** At the initial stage we can detect whether the patient is suffering from Breast Cancer or not.
* **Categorising Tumour as Benign or Malignant-** Tumours are mainly classified as Benign and Malignant so with the help of histopathologic breast cancer detection we can easily detect the tumours as harmful and harmless.
* **Analyze staging-** With the help of this project we can analyse the staging of the cancer and hence the patient can go for treatment without wasting time and before the staging proceeds.
* **Generating results based on analysis-** In this project we are tallying the images with the dataset to obtain the result.
* **Guiding patient through treatment stages-** After generating the results the patient will be guided that which type of treatment she must undergo.

**Proposed System**

Classification of tissues as Cancerous or Non Cancerous.

Accurate classification of cancerous tumors can prevent

patients undergoing unnecessary treatments.

Very large data sets are maintained, trained machine

learning models perform the processing.

Images are extracted , tallied with dataset and with the

use of AI the results classifying tumor are displayed to the user.

**PROBLEM STATEMENT**

**Problem Statement:**

Breast cancer is one of the most common cancers among women worldwide, ratio significantly states 1:8.

The major hurdles associated with this task involve:

(i) Need to transform oncogenic data for processing by computational methods.

(ii) Ability to extract interpretable and valid patterns from the processed, voluminous data.

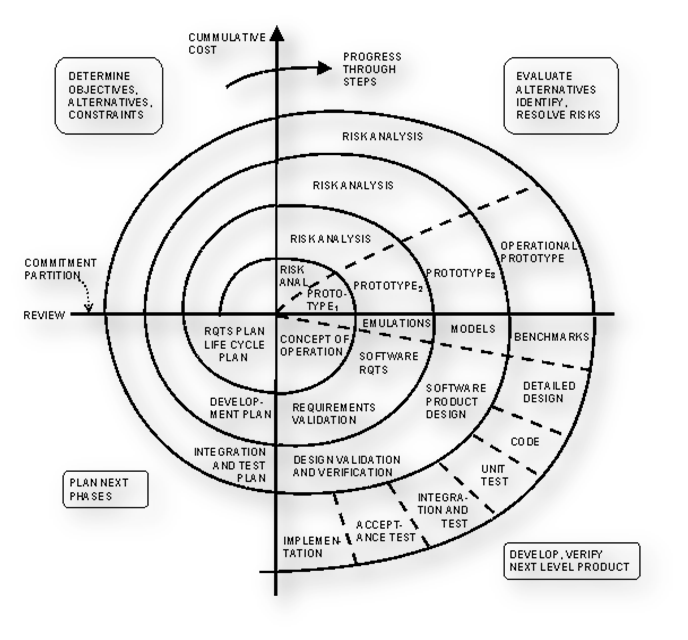
(iii)Improve prediction of cancer from diverse natured data with the extracted patterns.

**SYSTEM**

**ANALYSIS**

**System Design**

Spiral model is a combination of sequential and prototype model. This model is best used for large projects which involves continuous enhancements. There are specific activities which are done in one iteration (spiral) where the output is a small prototype of the large software. The same activities are then repeated for all the spirals till the entire software is build.



**Fig.1- Spiral Model**

**FEASIBILITY**

**STUDY**

Depending on the results of the initial test images the detection is now expanded to a more detailed feasibility study for “HISTOPATHOLOGIC BREAST CANCER DETECTION”.

Al projects are feasible given unlimited resources and infinite time. Unfortunately the development of computer-based system in many cases is more likely to be plagued by scarcity of resources and delivery date. Hence, we have made use the concept of reusability that is what Object Oriented Programming (OOP) is all about.

The feasibility report of the project holds the advantages and flexibility of the project. This is divided into three sections:-

1. Economic Feasibility

2. Technical Feasibility

3. Behavioral Feasibility

4. Operational Feasibility

**1. Economic Feasibility:-**

Economic analysis is the most frequently used method for evaluating the effectiveness of the candidate system. More commonly known as cost/benefit analysis, the procedure is to be determining the benefits and savings that are expected from a candidate and compare them with costs. If benefits outweigh costs, then the decision is made to design and implement the system.

A systems financial benefit must exceed the cost of developing that system. i.e. a new system being developed should be a good investment for the organization. Economic feasibility considers the following:

i. The cost to conduct a full system investigation.

ii. The cost of hardware and software for the class of application.

iii. The benefits in the form of reduced cost or fewer costly errors.

**2. Technical Feasibility:-**

Technical feasibility centers on the existing computer system (Hardware and Software etc.) and to what extend it support the proposed addition. In this project, all the necessary cautions have been taken care to make it technically feasible. Using a key the display of text/object is very fast. Also, the tools, operating system and programming language used in this localization process is compatible with the existing one. The technical needs of the system vary considerably but might include:-

i. The facility to produce outputs in a given time.

ii. Response time under certain conditions.

The project is technical feasible because of the availability of the required software hardware and technology. The changes can be made be made in the system as and when required.

**3. Behavioral Feasibility:-**

People are inherently resistant to change, and computers have been known to facilitate change. An estimate should be made of how strong a reaction the user staff is likely to have toward the development of a computerized system. Therefore it is understandable that the introduction of a candidate system requires special efforts to educate and train the staff. The software that is being developed is user friendly and easy to learn. In this way, the developed software is truly efficient and can work on any circumstances, tradition, locales. Behavioral study strives on ensuring that the equilibrium of the organization and status quo in the organization neither are nor disturbed and changes are readily accepted by the users. Thus these factors are considered for a Behavioral feasibility study:-

i. Need analysis

ii. Provide the user information pertaining to the preceding requirement.

**4. Operational Feasibility:-**

It determines how acceptable the software is within the organization. The evaluations must then determine the general attitude and skills. Such restriction of the job will be acceptable. To the users are enough to run the proposed budget, hence the system is supposed to the feasible regarding all except of feasibility. In operational Feasibility, we attempt to ensure that every user can access the system easily.

Operational feasibility of the project also exists because in today’s world most of the people are using the internet and are purchasing the products online. There is nothing complex in the system that cannot be used by people. It is socially accessible feasible as well because of its usefulness and easiness in getting information. Time feasibility also exists because it can be developed and implemented in the given time. As far as legal feasibility is concerned there is no such restriction faced by the system.

**Modules**

1. **Dataframe Loading**
2. **Data Validation System**
3. **Data Augmentation**
4. **High Level Feature Detection**
5. **Edge And Shape Detection Using RNN**
6. **Training System**
7. **Accuracy And Testing System**
8. **View Display Manager**
9. **Preference System**
10. **Reporting System**

**In this project, two members are involved and so, we have divided the whole project into two groups:-**

**The following modules are handled by Falguni Jiya:-**

* **Module 1** - DATAFRAME LOADING
* **Module 2** - DATA VALIDATION SYSTEM
* **Module 3** - DATA AUGMENTATION
* **Module 4** - HIGH LEVEL FEATURE DETECTION
* **Module 5** - EDGE & SHAPE DETECTION USING RNN

**The following modules are handled by Ayush Tiwari:-**

* **Module 6 –** TRAINING SYSTEM
* **Module 7 –** ACCURACY & TESTING SYSTEM
* **Module 8 –** VIEW DISPLAY MANAGER
* **Module 9 –**  PREFERENCE SYSTEM
* **Module 10 –** REPORTING SYSTEM

**Module Description**

**1. Module 1** - DATAFRAME LOADING:

Loads the whole chunk of data through python code.

**2. Module 2** - DATA VALIDATION SYSTEM:

Validates if the dataset are loaded & are complete.

**3. Module 3** - DATA AUGMENTATION:

**A**llows the system to allow new images to be added.

**4. Module 4** - HIGH LEVEL FEATURE DETECTION:

Detection of image feature important for image recognition.

**5. Module 5** - EDGE & SHAPE DETECTION USING RNN:

Require for the system to understand the shape of cell in tissues.

**6. Module 6 –** TRAINING SYSTEM:

It will use the transfer learning approach, which will be trained in 12

epoch or more & take most of the time.

**7. Module 7 –** ACCURACY & TESTING SYSTEM:

Checks the AI trained model performance.

**8. Module 8 –** VIEW DISPLAY MANAGER:

Interface for the application to check the program.

**9. Module 9 –** PREFERENCE SYSTEM:

Setting for the project.

10. **Module 10 –** REPORTING SYSTEM:

Generate prediction result.

**SOFTWARE**

**REQUIREMENT SPECIFICATION**

**Gantt Chart**

A Gantt chart is popular type of chart that illustrates a project schedule. Gantt Chart illustrates the start and finish dates of the terminal elements and summary elements of a project. Terminal element and summary comprise the work breakdown structure of the project.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Task** | **4Jan-30Jan** | **31Jan-9Feb** | **10Feb-12Mar** | **13Mar-16Apr** | **17Apr-22Apr** | **23Apr-28Apr** |
| **Develop project proposal** | 27 days |  |  |  |  |  |
| **Analysis** |  | 10 days |  |  |  |  |
| **Designing** |  |  | 30 days |  |  |  |
| **Coding** |  |  |  | 34days |  |  |
| **Unit Testing** |  |  |  |  | 5 days |  |
| **Implementation** |  |  |  |  |  | 5 days |

**Functional Requirements:**

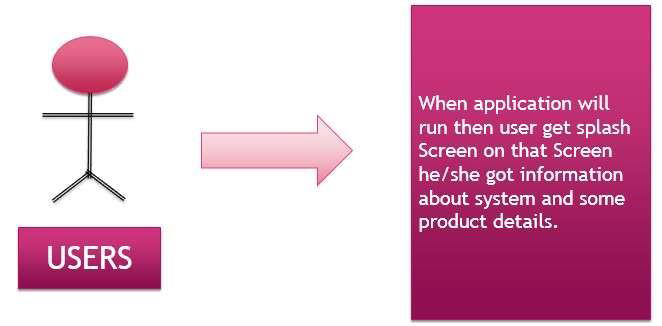
|  |  |  |  |
| --- | --- | --- | --- |
| SR. NO | REQ. NO. | PRIORITY | REQUIREMENTS |
| 1. | R1.1 | C | Installation of application |
| 2. | R1.2 | R | Start application |
| 3. | R1.3 | R | Input through images |
| 4. | R1.4 | R | Tallying of images |
| 5. | R1.5 | R | prediction |
| 6. | R1.6 | R | Output shown on screen |

**Non Functional Requirements:**

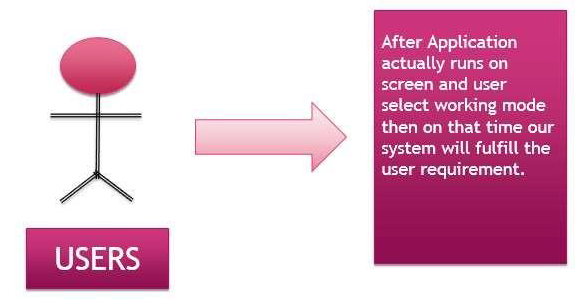
|  |  |
| --- | --- |
| CATEGORY | NON FUNCTIONAL REQUIREMENTS |
| USABILITY | Users must be able to install the app |
| RELIABILITY | App permission would be important and will not use any private data of users |
| PERFORMANCE | The will able to give good perform on most of the OS |
| SUPPORTABILITY | This software supports all windows above windows 7 |
| IMPLEMENTATION | All users must access the software by opening it in windows device |
| OPERATION | All the operations performed by user settings in the software |
| LEGAL | The software should be used as required by local law |

**ROLE DESCRIPTION**

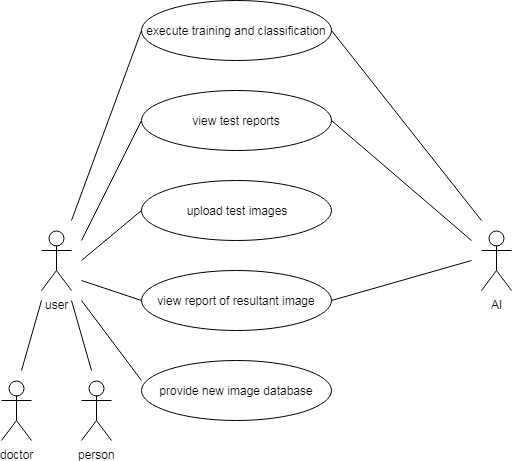
**Stimulus/Response Sequence**



**Description & Priority**



**8. Use Case Diagram:**



**9. Software Requirement:**

**Client Side**

● Web Browser (Google Chrome, Firefox, IE9 or above)

**Server Side**

● Web Browser (Google Chrome, Firefox, IE9 or above)

● Windows 7 or above / Linux

● Pycharm IDE 2016

● SQliteManager

● Redis,Node.js

● Python 3.5, Django

**10. Hardware Requirements:**

CLIENT SIDE

|  |  |
| --- | --- |
| Processor | Dual Core or above |
| RAM | 1 GB |
| Disk space | 500 GB |
| Monitor | 15” |
| Others | Keyboard, mouse, Internet Connection |

SERVER SIDE

|  |  |
| --- | --- |
| Processor | Dual Core or above |
| RAM | 1 GB |
| Disk space | 500 GB |
| Monitor | 15” |
| Others | Keyboard, mouse, Internet Connection |

**Document convention**

•**Font Family**

Times New Roman

Georgia

•**Font Size**

Heading – 28

Paragraph - 14

•**Font Color**

Content-black

**THE OPERATING SYSTEM:**

The operating system used is windows 10, because it is the latest windows operating system. The software can also run on windows 8 or 8.1 as well as Linux system.

**PYTHON:**

Python is an interpreted, high-level, general-purpose programming language. Created by Guido Van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales.

**TENSORFLOW:**

TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks.

**KERAS:**

Keras is an open-source neural-network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, Theano, or PlaidML. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible.

**OPENCV:**

OpenCV is a library of programming functions mainly aimed at real-time computer vision. Originally developed by Intel, it was later supported by Willow Garage then Itseez. The library is cross-platform and free for use under the open-source BSD license

**TKINTER:.**

Tkinter is a Python binding to the Tk GUI toolkit. It is the standard Python interface to the Tk GUI toolkit, and is Python's de facto standard GUI. Tkinter is included with standard Linux, Microsoft Windows and Mac OS X installs of Python. The name Tkinter comes from Tk interface.

**NUMPY:**

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

**PILLOW:**

Python Imaging Library is a free library for the Python programming language that adds support for opening, manipulating, and saving many different image file formats. It is available for Windows, Mac OS X and Linux.

**DESIGN IMPLEMENTATION & CONSTRAINTS**

**Constraint 1:** Minimum 1 GB of hard disk on user is required to setup the software.

**Constraint 2:** System with proper internet connection is required

**Constraint 3:** All libraries are required to be installed on user system.

**SECURITY REQUIREMENTS**

There is only one user of this software , he/she can be doctor or any other normal person with good knowledge of internet and computer.

The doctor or the other person will take the mamographic image and upload it to the software. After tallying the mamographic image with the data set the HISTOPATHOLOGIC BREAST CANCER DETECTION software will generate the results and show whether the cell is cancerous or non cancerous.

**INTENDED AUDIENCE AND READING SUGGESTIONS**

**The MULTISITE WEB CRAWLER FOR PRODUCT RECOMMENDATION SYSTEM is**

intended to be used for everyone who want to plan their learning in proper manner where they can choose various crawler in order to get the most efficient product recommendations.

There is different Audience of this SRS document. Some are-

**Ø Developer-** Developer uses this SRS for maintainability & modification of this web project. They also use this document to understand the process flow of the system**.**

**Ø Client-** Client uses this SRS document to verify the pre-defined requirement. They can also suggest some further requirement in future according as per their need.

**Ø Tester-** Tester perform black box testing or white box testing according to predefined requisites provided in SRS.

**Ø System Analyst-** Analyst uses this document for system testing and they use it for analysis of the proposed system as required by client.

**PROJECT CATEGORY**

**1.** This project can be categorized under machine learning

**2.** MULTISITE WEB CRAWLER FOR PRODUCT RECOMMENDATION SYSTEM is purely a software.

**3.** It is an online project which is meant to run on internet & to services to its user’s.

**4.** Backend used relational database management system SQlite, Python having animations developed in Pycharm .

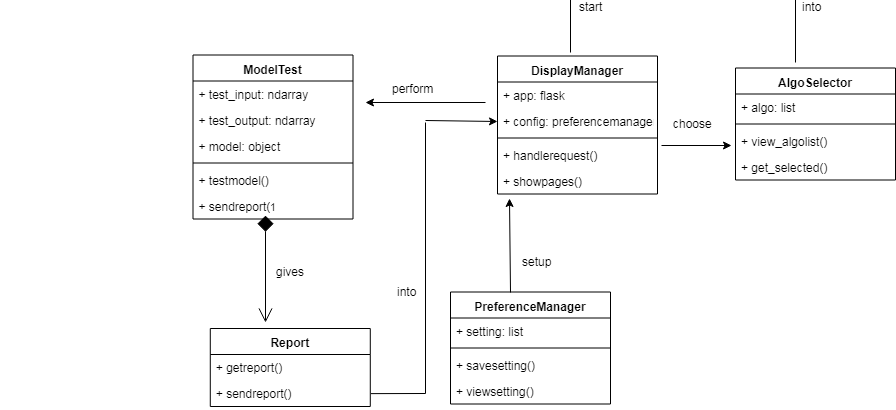
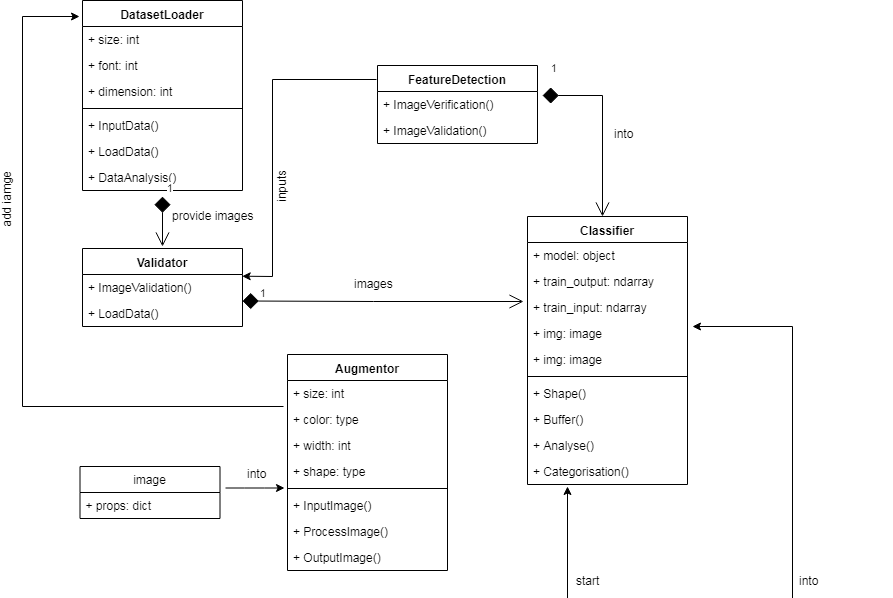
**5.** Design of the website has been developed in QT designer.

**6.** Language used is python.

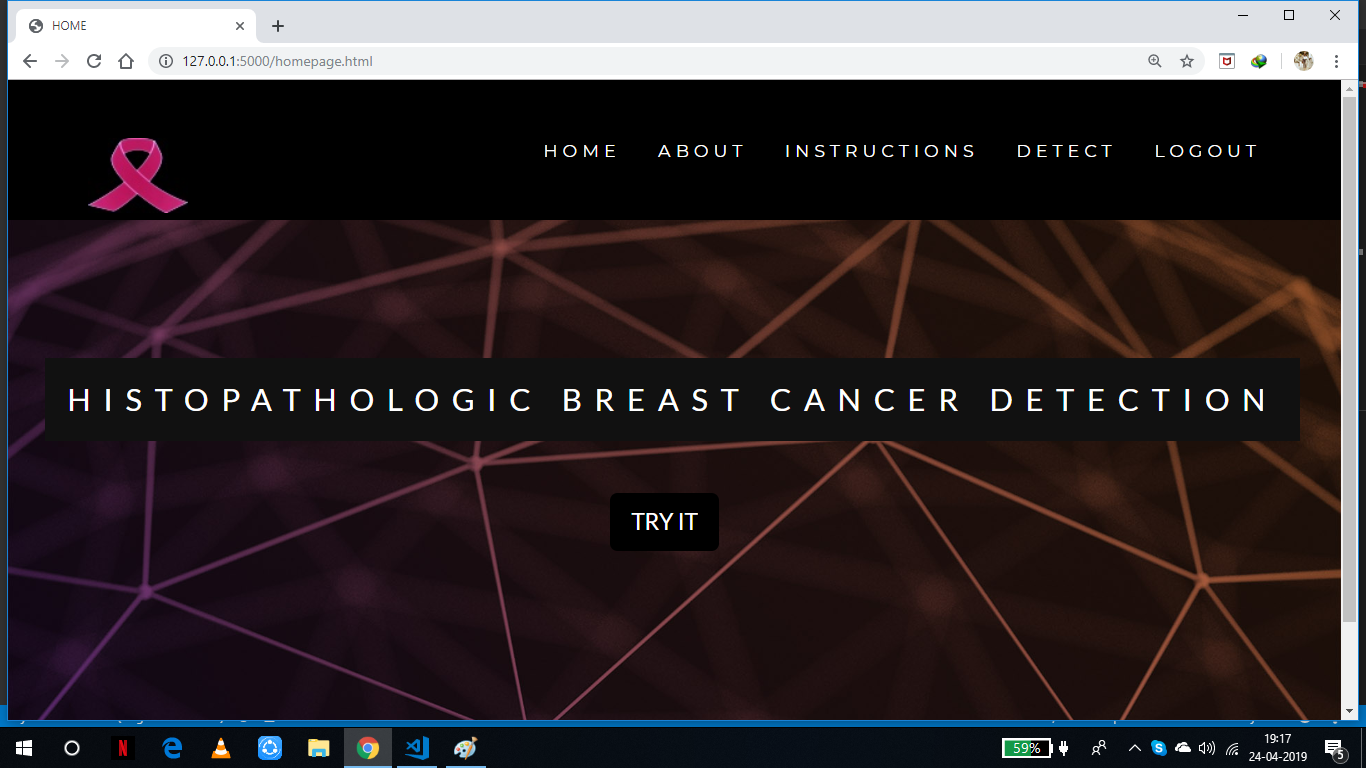
**7.**Use of machine learning in order to develop recommendation system.

**HIGH LEVEL DESIGN**

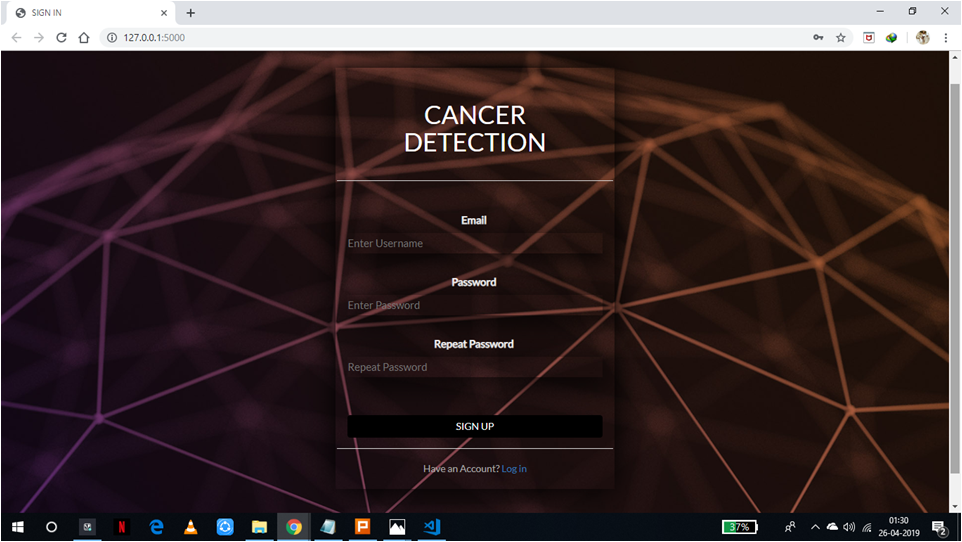
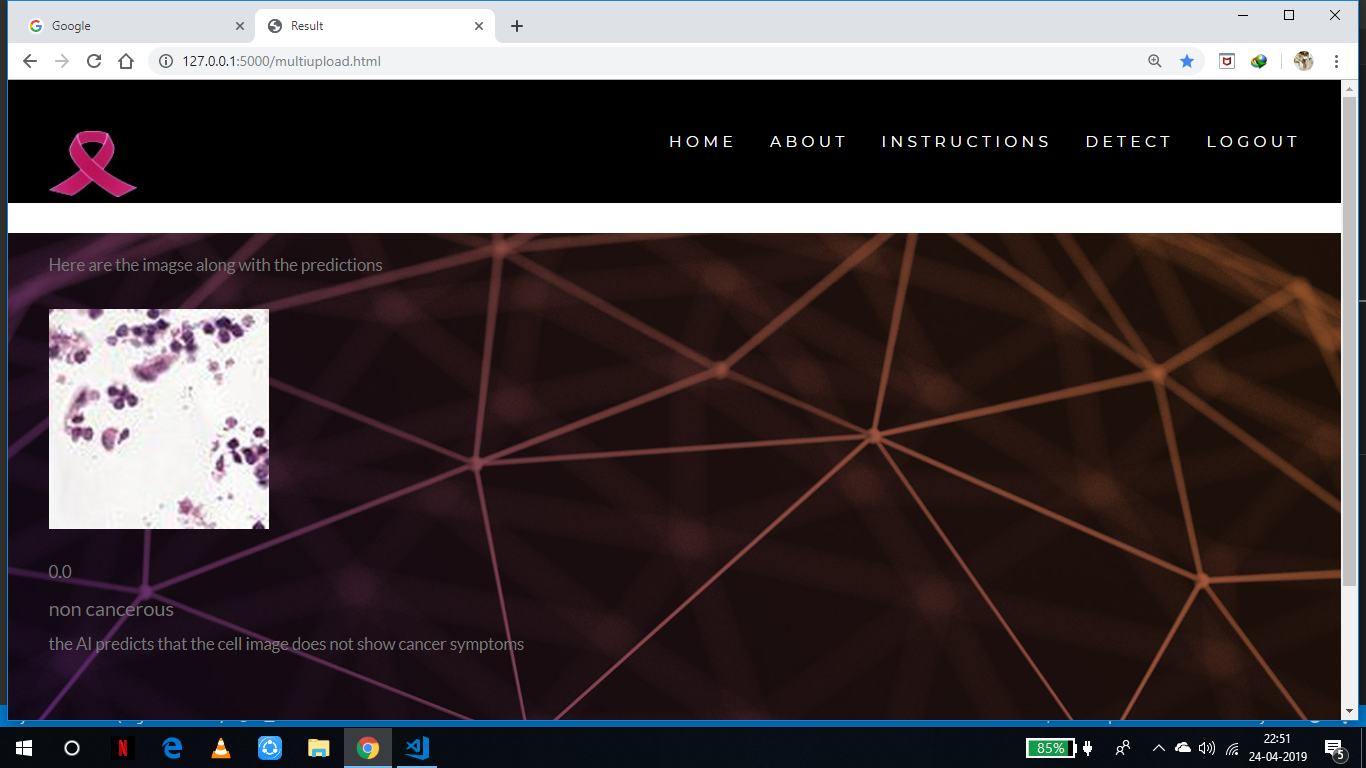
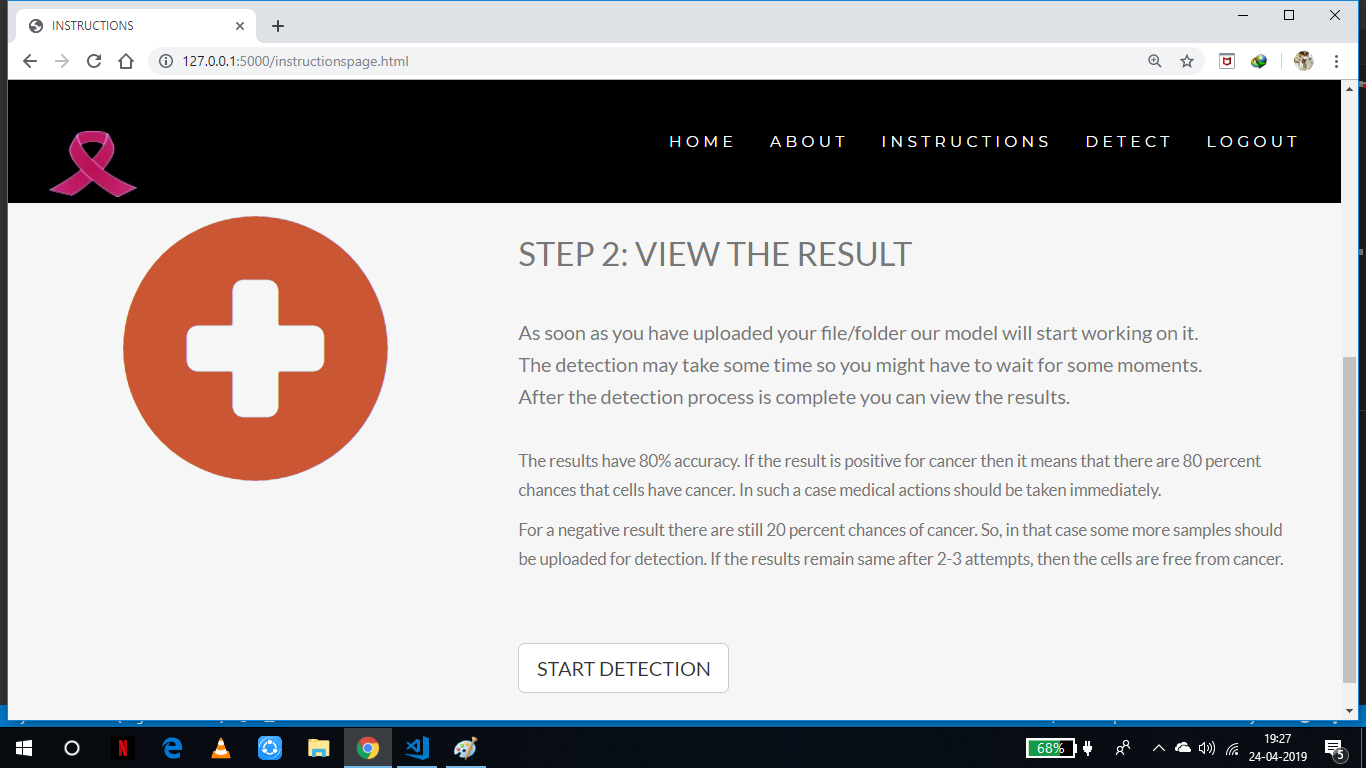
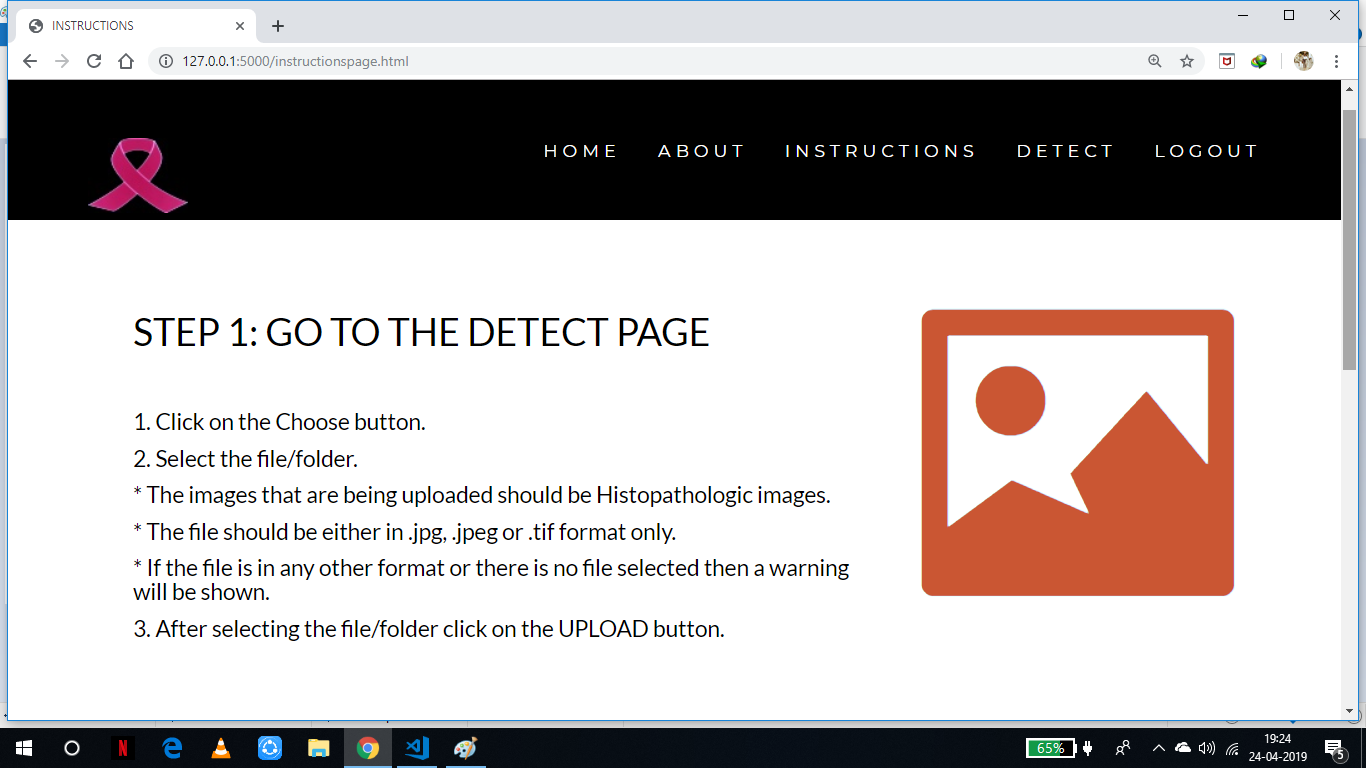
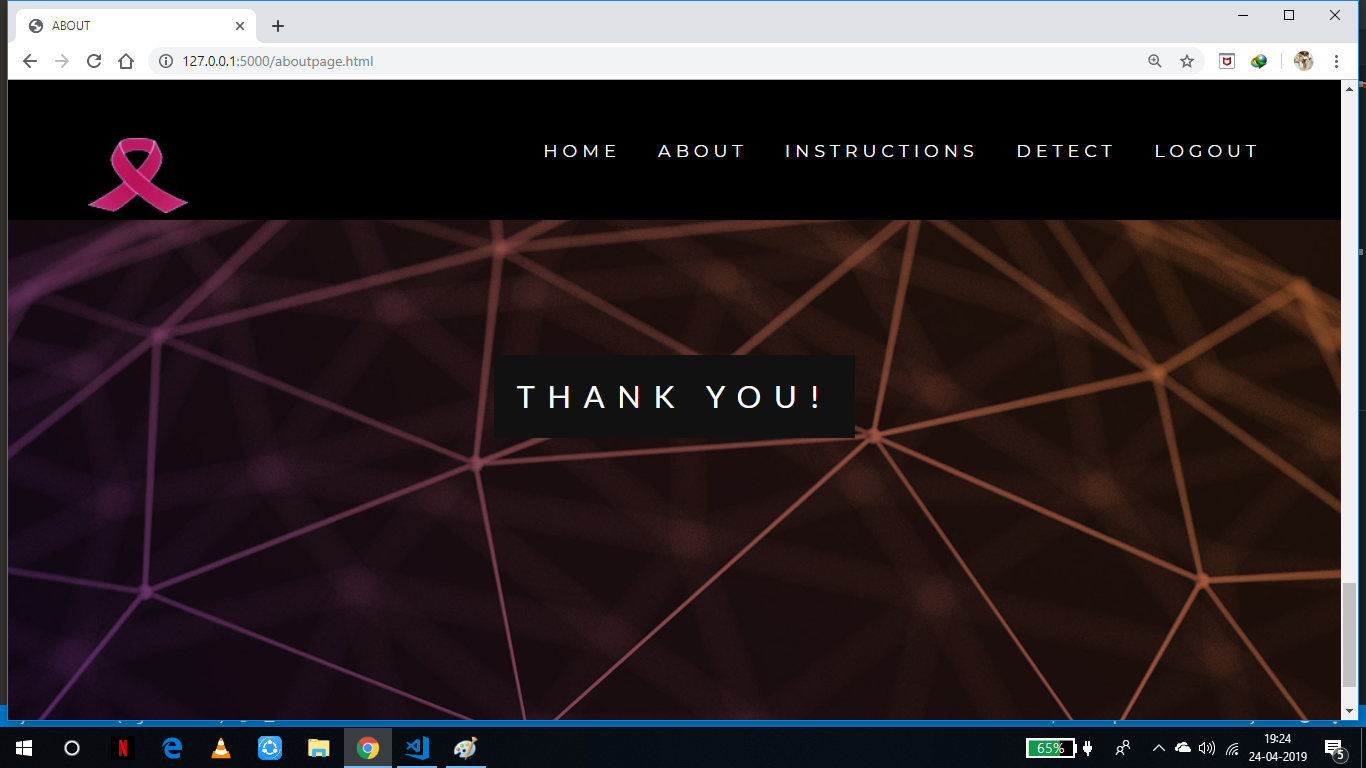
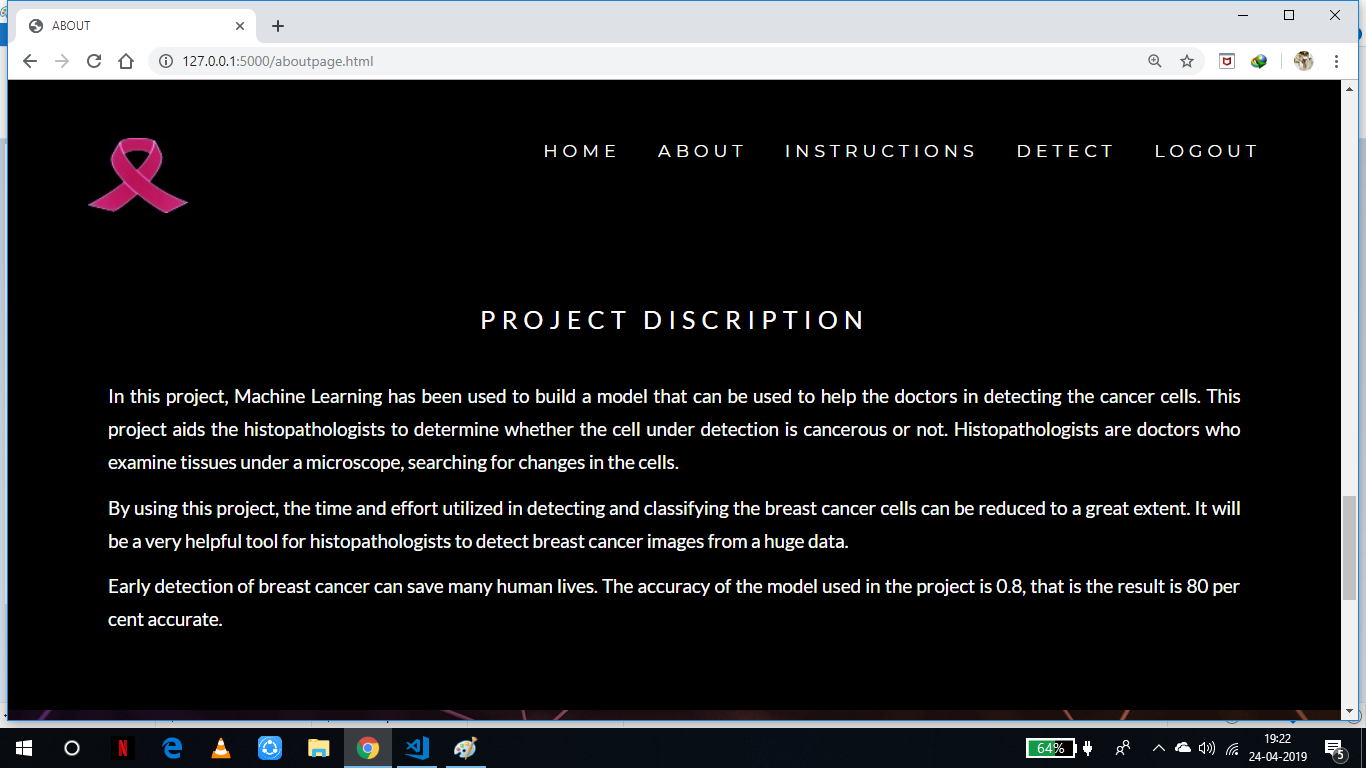
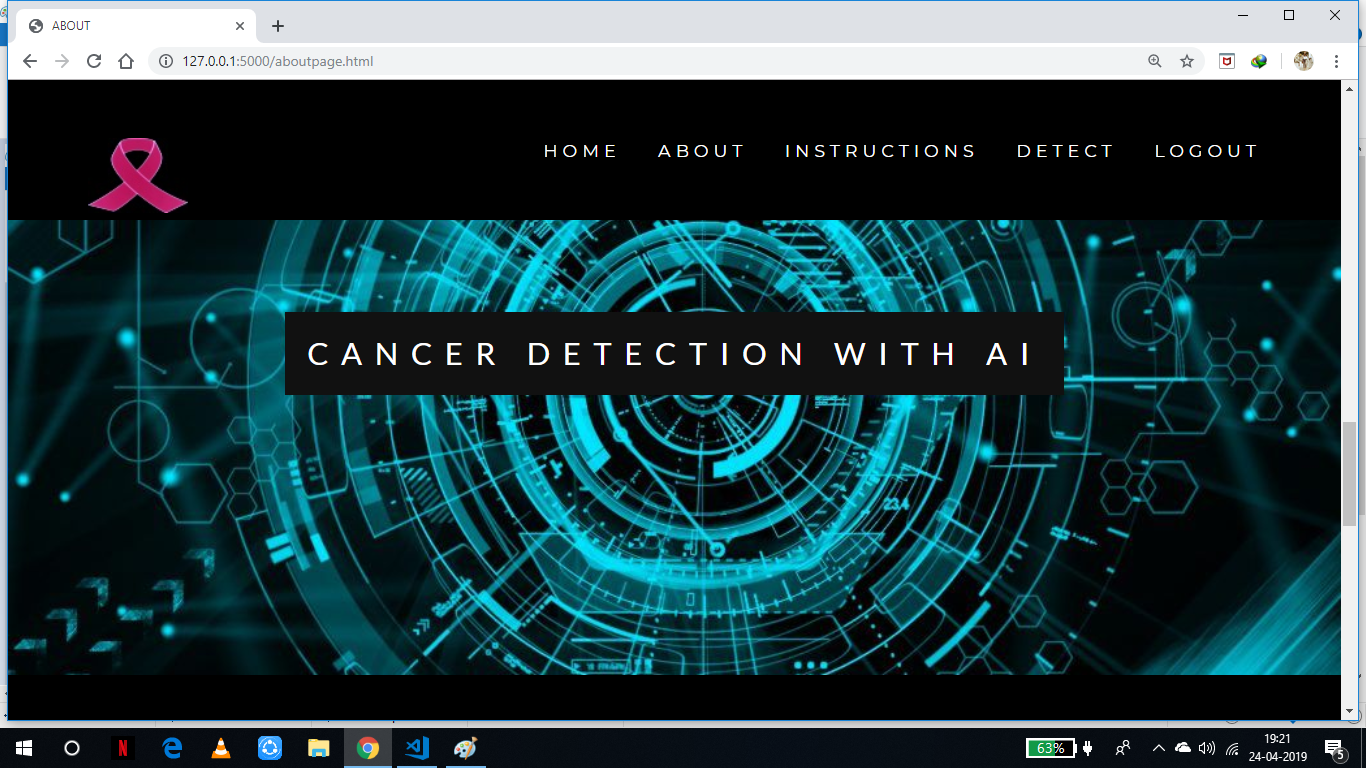
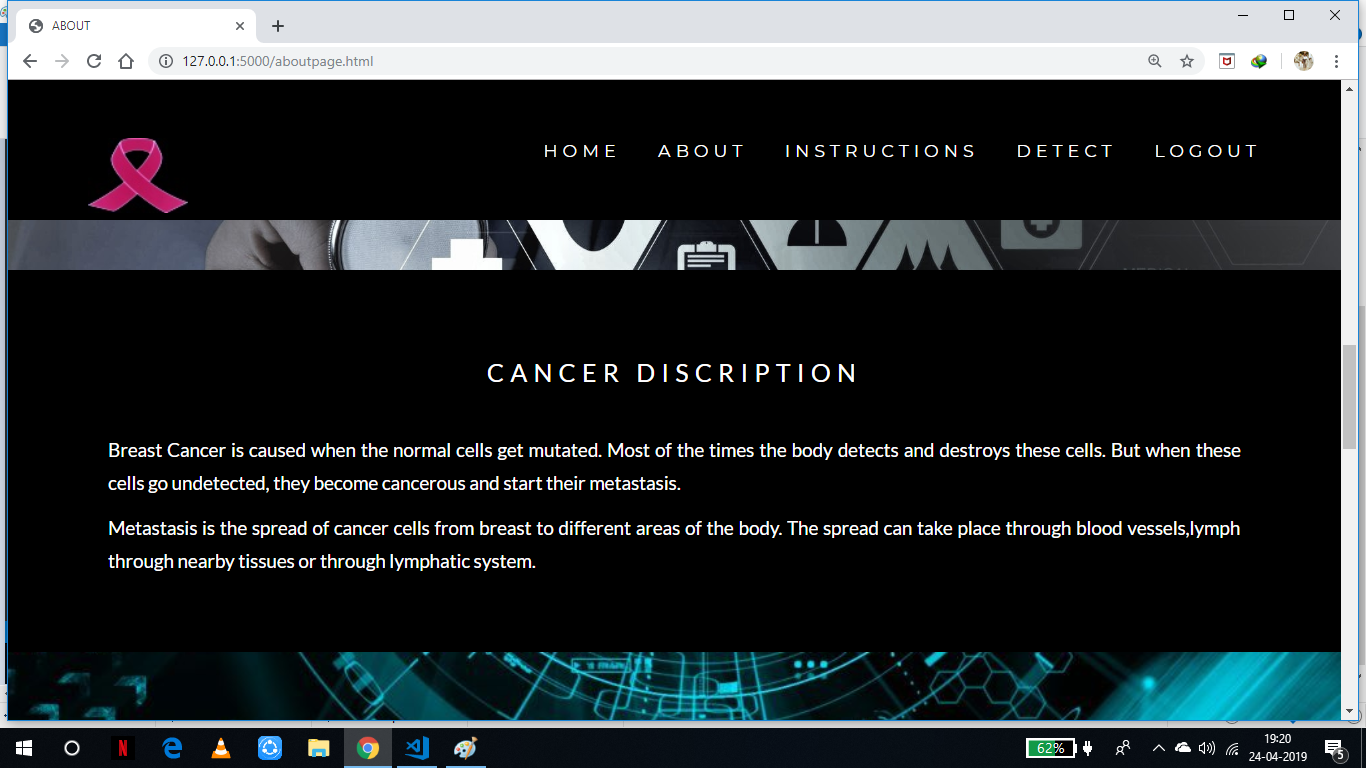
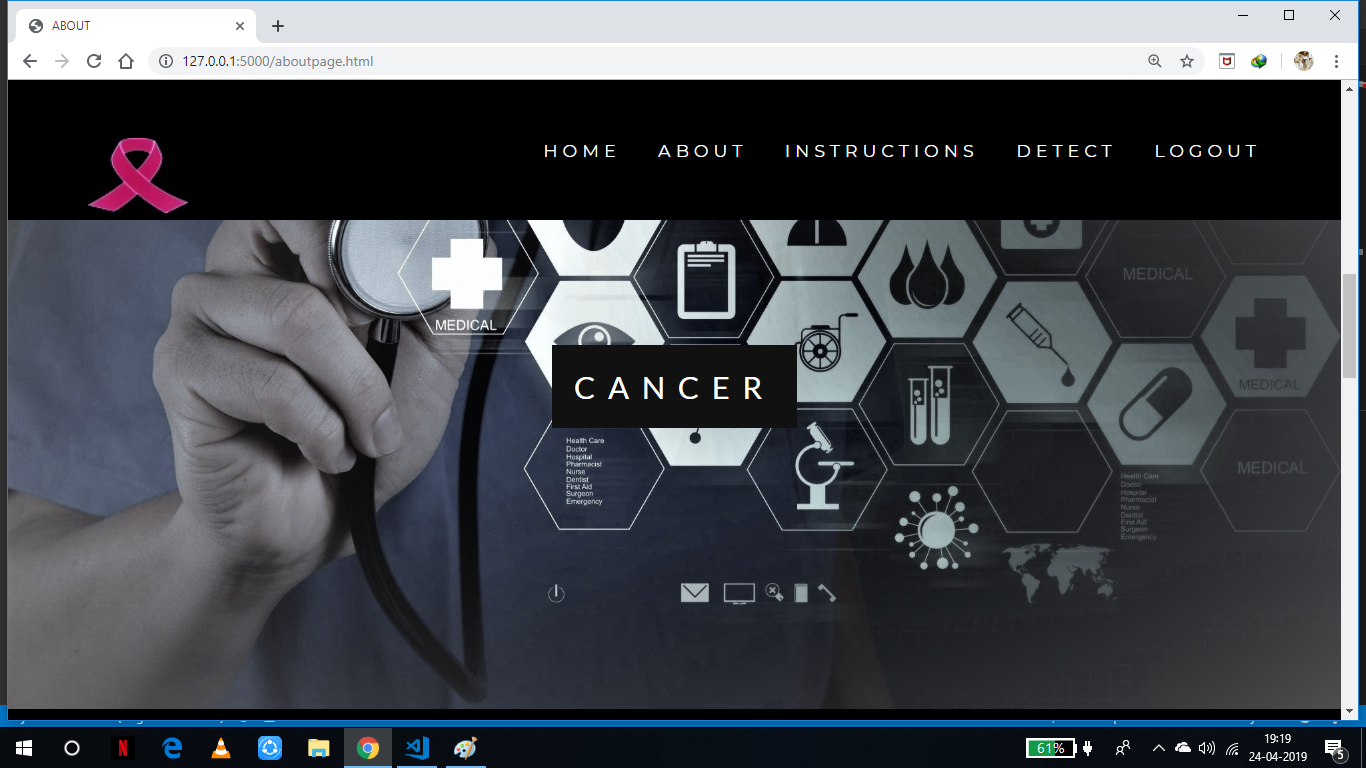
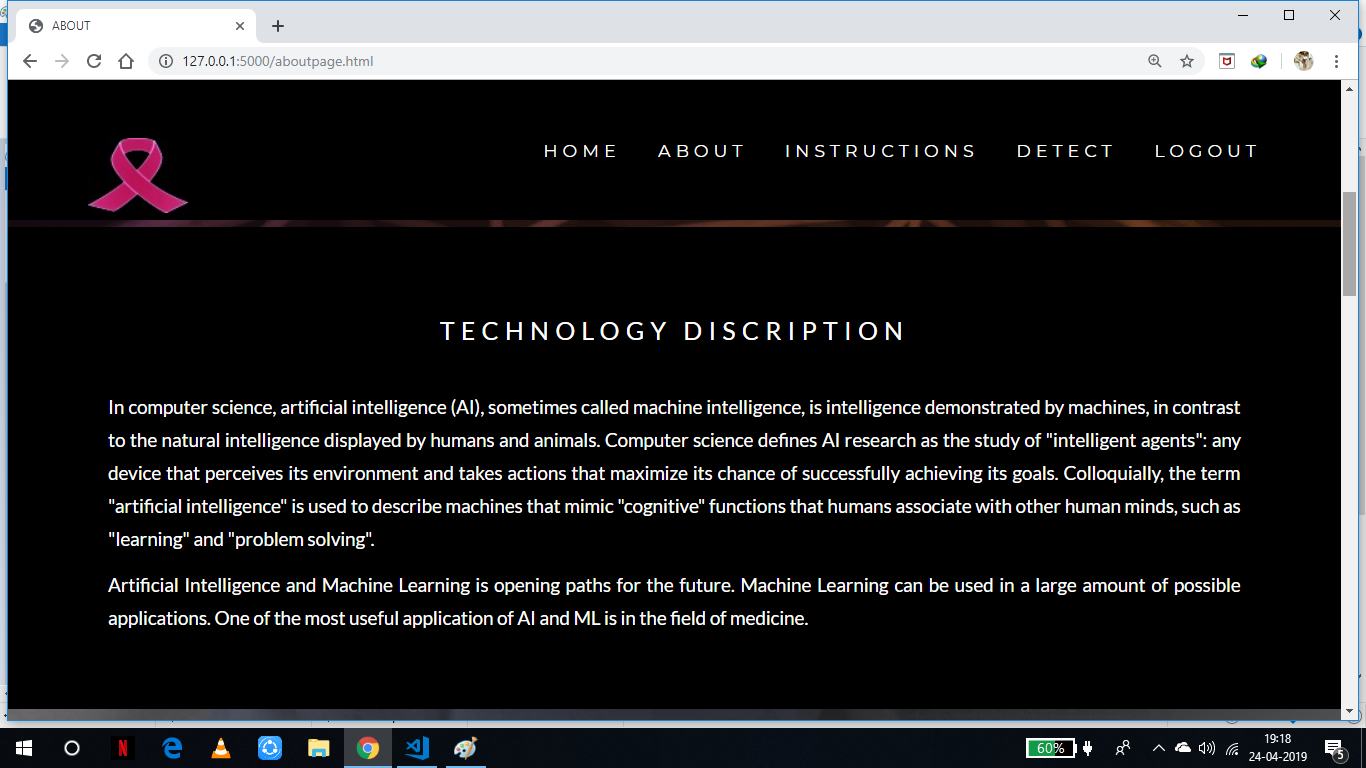
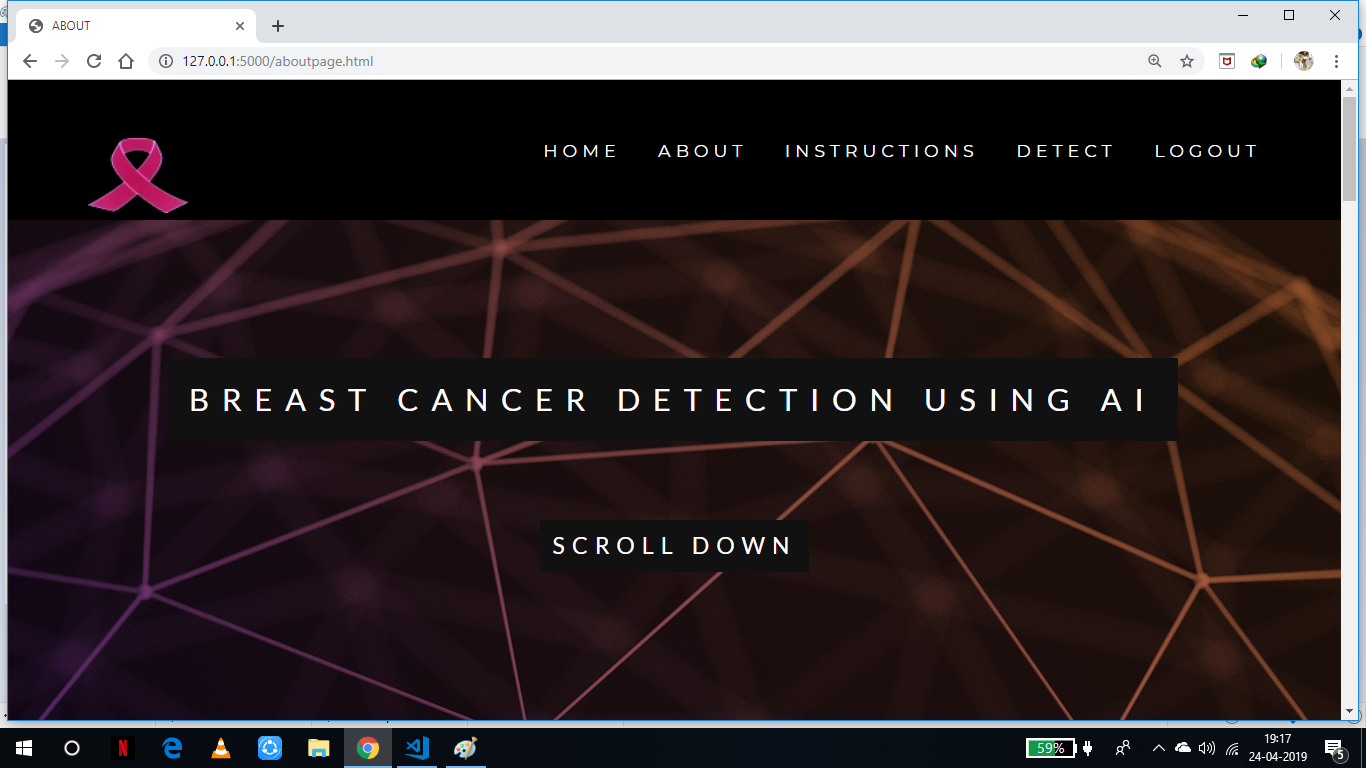
**UML Diagram**



**SCREENSHOTS OF SOFTWARE**



**Home Page**



**TESTING IMPLEMENTATION AND FUTURE SCOPE**

**TEST CASES**