BREAST CANCER DIAGNOSIS USING MACHINE LEARNING:

A SIMPLE EXPERT SYSTEM

Justin Rey Ledesma

Angel Edith Saludes

Jan Angelo Gorres

BSDS-3A

January 11, 2021

SUMMARY SHEET

The expert system of breast cancer diagnosis helps the doctors to see the results of diagnosis faster and more accurate decision. However, the experienced staffs need to guide the inexperience staff to improve their handling of the diagnosis. This system focuses on specific areas of disease in order to improve the accuracy of results. The researchers used Random Forest classifier to get a more accurate and stable prediction for our data.

ACKNOWLEDGEMENT

We would like to acknowledge and praise God, the almighty, who has granted countless blessings and knowledge to accomplish this task.

We would also like to record our appreciation to all the people involved in this project. First of all, to our instructor, Engr. Dave Charity Gambuta for giving us the opportunity to perform this task and for the guidance until we complete this project. In addition, we are also grateful to our family, friends and specially to our BSDS-3A classmates for all the help and support.

LIST OF FIGURES

Figures **Pages**

ABSTRACT

Breast cancer is a type of cancer that is widespread among women worldwide, with nearly 1.7 million new cases diagnosed in 2012, representing about 25 percent of all cancers in women. In this simple expert system in breast cancer diagnosis, the aim is to be able to early detect breast cancer and classify if it’s malignant or benign without doing long processes and various tests as machine learning do the work. Machine learning as a tool in the detection of breast cancer is widely being used in many studies around the world, may it be in using images, inputs records and etc. The attempt to really apply supervised machine learning in the field of medicine, specifically in the BC diagnosis is not new today. The performance of the proposed method was based on accuracy, recall and F-1 score. With the help of our GUI programming techniques and machine learning algorithms we come up with an accuracy of 95 % using random forests classifier.

\

1.0 INTRODUCTION

Breast cancer is considered to be one of the most common cancer in the world. In some cases, it can occur in men but it is most common among women. Approximately 30 % of new women cancer diagnoses will be breast cancer in 2020 as estimated by the World Health Organization. The Mayo Clinic Organization also stated that Doctors know that breast cancer occurs when some breast cells begin to grow abnormally. When these breast cells grow out of control, that’s when the disease comes in. These cells divide more rapidly than healthy cells do and continue to accumulate, forming a lump or mass. Cells may spread (metastasize) through your breast to your lymph nodes or to other parts of your body. These lumps can either be identified and evaluated as benign (non-cancerous) or malignant (cancerous) through physical examination performed by a doctor.

Over the years, several studies related to the diagnosis of breast cancer using algorithms has been conducted in the attempt to automate detection of the breast cancer. As technological advancement advances continuously and rapidly this simple expert system would like to apply it by mimicking a doctor’s clinical decision and by making it possible to identify and diagnose the breast lump as cancerous and non-cancerous through the use of a supervised learning approach, machine learning.

1.2 STATE OF THE ART

The Expert system of breast cancer diagnosis doesn't require high memory capacity to be fully functional, because it just needs the data given from the patient for the diagnosis to identify whether tumor is benign (noncancerous) or malignant (cancerous). However, in time, large RAM capacity and computing power will be necessary, as this ES will be improved in time. An expert system's information is obtained from expert sources and coded in a form suitable for the system to use in its suggestion for reasoning process. The expert system information should be obtained from professional or other sources of expertise. The inexperience staffs need the guide from the experienced staffs to improve their skill in handling the diagnosis. It also to reduce the time required to come to a decision particularly in an emergency case.

1.3 SIGNIFICANCE

Over the years, mammography is the process used in examining human breasts for screening and for cancer diagnosis thus having a goal of an early detection but it does not guarantee a 100 % accuracy. American Cancer Society states that limitations of mammograms are showing false-negative and false-positive results, by this, follow-up check-ups and other tests are necessary before concluding the diagnosis. To lessen that, this expert system will benefit and merit the following:

Patients. Patients often goes through a lot of waiting time, long processes of results, various tests before, this expert system may serve as a tool in optimizing patients' waiting time and providing a more effective health care for the patients with a more accurate and quality results.

Doctors. Doctors performs clinical exams, mammography and ultrasounds before coming up with a diagnosis. This expert system may serve as an aid to doctors for them to make better and faster diagnoses and a one step ahead decisions for their patients without going through a lot of tests avoiding wrong medication.

Health Care Systems. Data and Analytics is the top four biggest issues in health care nowadays**.** This expert system may serve as a tool to improve the health care system of the country to find better treatment options to become more efficient and reliable while dealing with massive amounts of data.

1.4 OBJECTIVES

1. Identify and describe the clinical presentations of breast cancer using machine learning.

2. Identify whether condition is benign or malignant.

3. To obtain diagnosis faster, using machine learning.

2.0 REVIEW OF LITERATURE

This chapter includes ideas and finished studies related to breast cancer diagnosis using machine learning algorithms that are relevant and to help in familiarizing information about the proposed expert system.

According to Wenbin Yuer et.al, (2018), breast cancer being one of the most common cancers especially on women worldwide represents the majority of new cancer cases and cancer-related deaths as per global statistics thus making it as one of the top health problems nowadays. With this, a **"Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis"** was conducted to review machine learning techniques and how they can be applied in the breast cancer diagnosis and prognosis, the research study concluded that "Machine Learning techniques have shown their remarkable ability to improve classification and prediction accuracy." after various methods used by the researchers during the process of conducting their study.

The initial study encountered was by Cook & Fox on **Artificial Intelligence in Breast Cancer Research**, where mammographic image analysis was investigated via a decision table to characterize all the parameters and potential in 41 rules that were produced, all centered on masses and lesions.

Also, it applied (Artificial Neural Network (ANN) on mammography for decision making in the diagnosis of breast cancer. A network that used image features performs well in distinguishing between malignant and benign lesions. Predicted breast cancer malignancy tumor using an ANN on a retrospective set of data of patients scheduled for biopsy, breast biopsy decisions. Results of biopsies were taken as the truth in diagnosis of the malignancies.

Breast manual examinations takes periods of waiting with false-positive results, this lead researchers, data miners, and other related field to use technological advancement in the health sector. As per Hiba Asri et.al, (2016), a study entitled: **"Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis"** conducted a performance comparison between different machine learning algorithms: Support Vector Machine (SVM), Decision Tree, Naive Bayes and k Nearest Neighbors (k-NN) to assess the effectiveness and correctness of these algorithms in the prediction and classification of breast cancer, having SVM 97.13% of accuracy in predicting the cancer.

Xin Yao et al. 1999 on **Breast Cancer Diagnosis using Neural Networks** has attempted to implement neural network for breast cancer diagnosis. Negative correlation training algorithm was used to decompose a problem automatically and solve them.

In this article the author has discussed two approaches such as evolutionary approach and ensemble approach, in which evolutionary approach can be used to design compact neural network automatically. The ensemble approach was aimed to tackle large problems but it was in progress.

In addition, a lot of team performed researches to study the application of machine learning in the breast cancer diagnosis. According to Konstantina Kourou et.al, (2015) in their study on **Machine learning applications in cancer prognosis and prediction**, “The main objective of machine learning techniques is to produce a model which can be used to perform classification, prediction, estimation or any other similar task.” The researchers also used a variety of techniques, this includes Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs). Classification problem is a procedure under supervised learning, for example the researchers collected health records of breast cancer patients and will try to classify if it’s malignant or not, machine learning can work as the probability that it is malignant or not is Yes = 1 and No = 0. This study concluded that machine learning can provide promising tools for inference in the cancer domain.

3.0 THEORETICAL FRAMEWORK

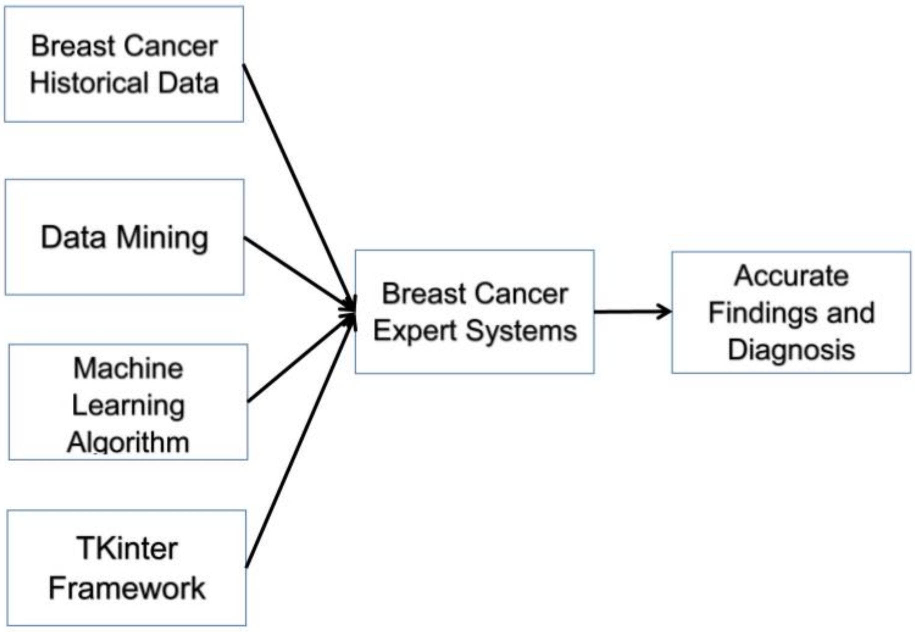

Breast Cancer Historical Data
Data Mining
Machine Learning Algorithm
TKinter Framework
Breast Cancer Expert Systems
Accurate Findings and Diagnosis

Figure 3.1 Theoretical Framework

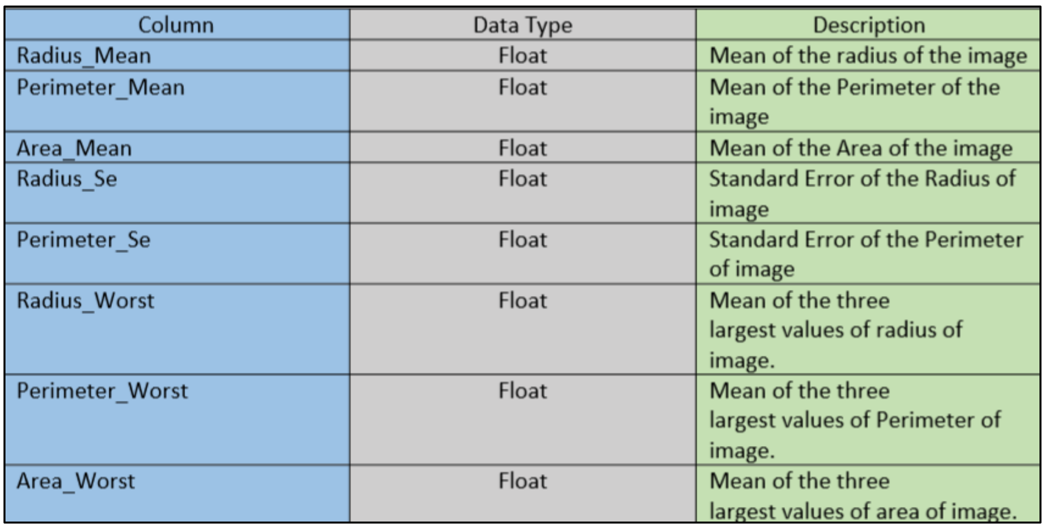
The theoretical framework shows the path of the expert system study and basis it firmly in theoretical constructs. The efficacy of this expert system, lies on these four major steps and algorithms. With these four, it's affirmative that the expert system could do its task efficiently and effectively.

4.0 METHODOLOGY

In this chapter the methodology used was presented, that includes the data requirements, project architecture and design, project implementation and software testing.

4.1 DATA REQUIREMENTS

The following table is the data requirements of the expert system for the Breast Cancer Diagnosis:



Column	Data Type	Description
Radius_Mean	Float	Mean of the radius of the image
Perimeter_Mean	Float	Mean of the Perimeter of the image
Area_Mean	Float	Mean of the Area of the image
Radius_Se	Float	Standard Error of the Radius of image
Perimeter_Se	Float	Standard Error of the Perimeter of image
Radius_Worst	Float	Mean of the three largest values of radius of image.
Perimeter_Worst	Float	Mean of the three largest values of Perimeter of image.
Area_Worst	Float	Mean of the three largest values of area of image.

Figure 4.1 Data Requirements

4.2 PROJECT ARCHITECTURE AND DESIGN

The following sets of diagrams illustrates the concepts and detailed process of the expert system. This includes, machine learning flow chart, flow chart diagram, use case diagram and the context diagram.

- Machine Learning Flowchart

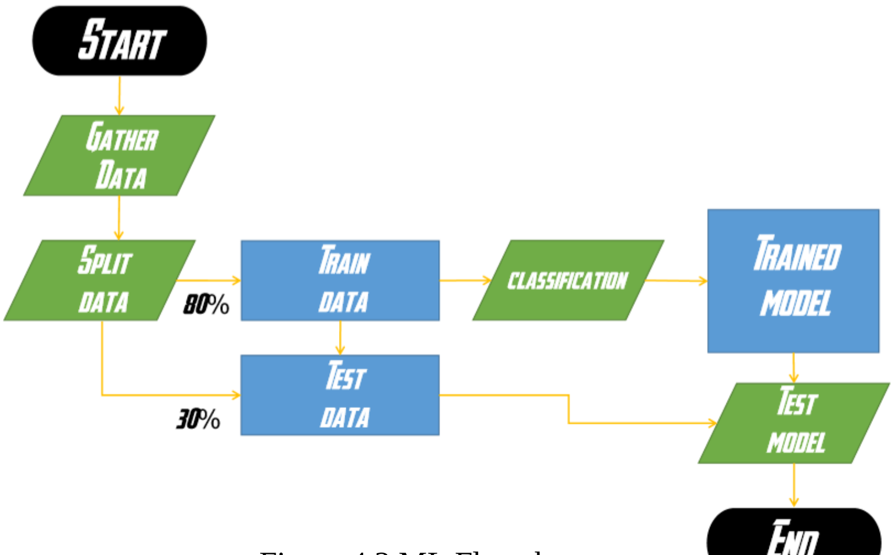


Figure 4.2 ML Flowchart

- Random Forest Classifier Flowchart Diagram

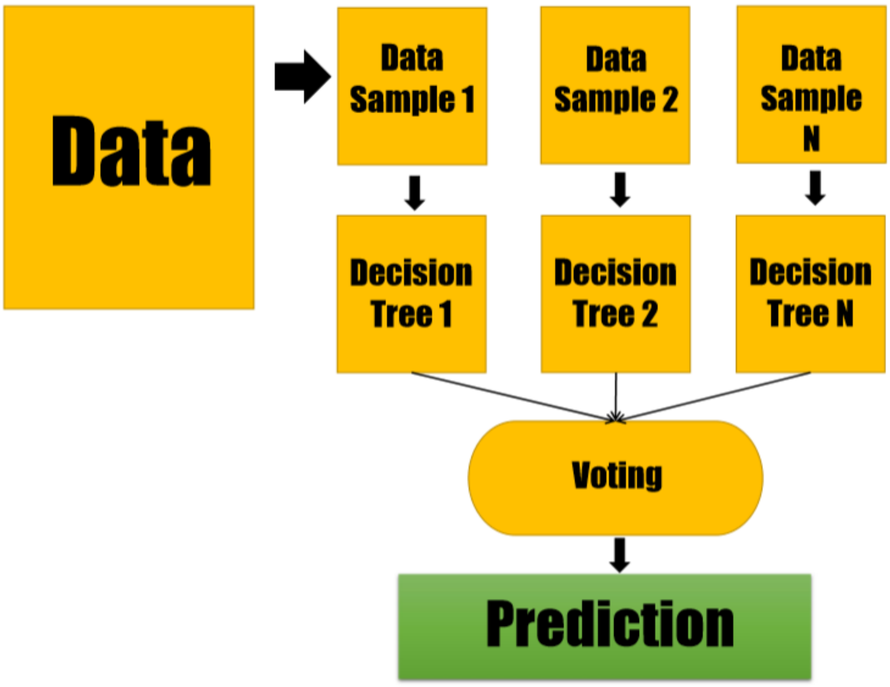
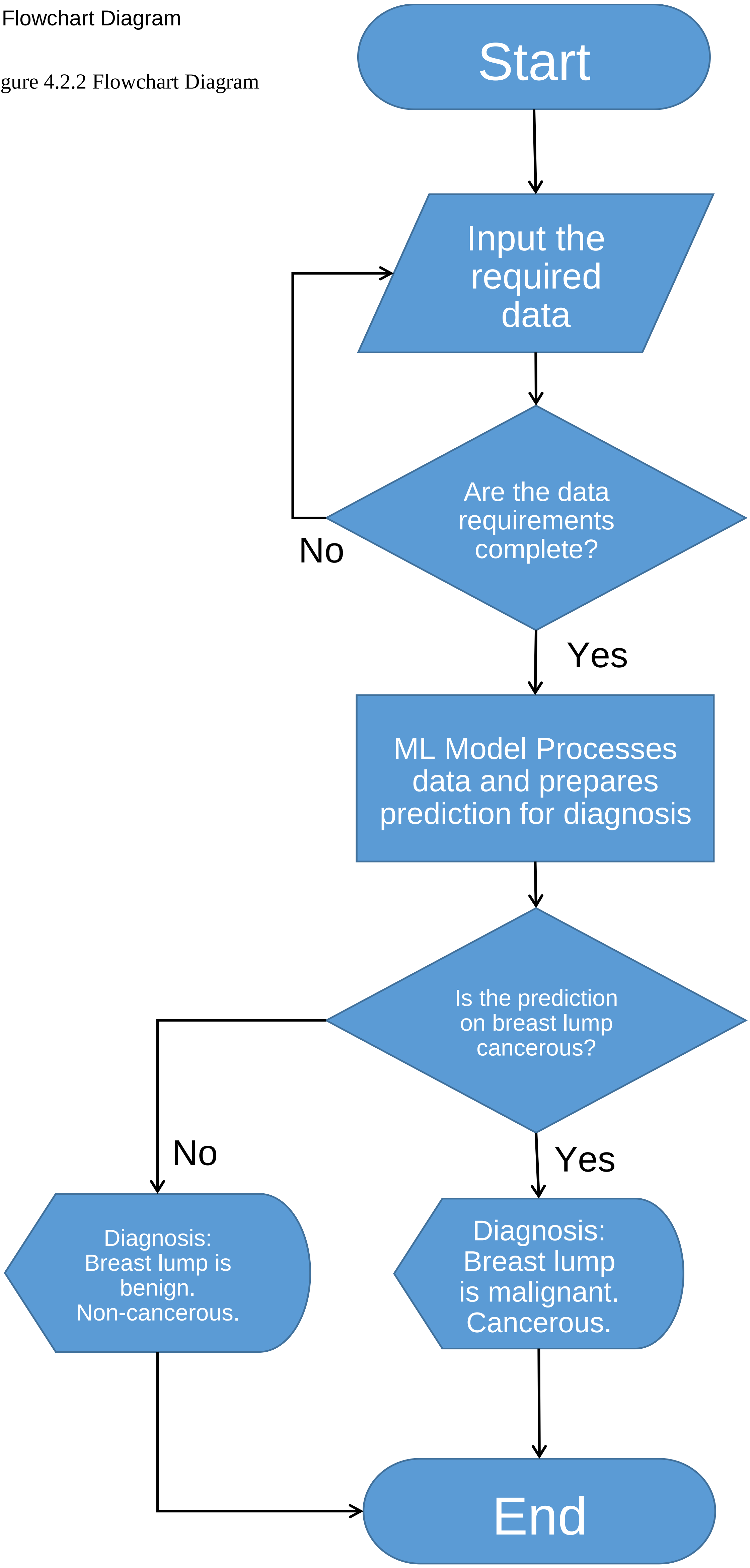


Figure 4.2.1 RFC Flowchart Diagram

- Flowchart Diagram

Figure 4.2.2 Flowchart Diagram


Start
Input the required data
Are the data requirements complete?
No
Yes
ML Model Processes data and prepares prediction for diagnosis
Is the prediction on breast lump cancerous?
No
Yes
Diagnosis: Breast lump is benign. Non-cancerous.
Diagnosis: Breast lump is malignant. Cancerous.
End

- Use Case Diagram

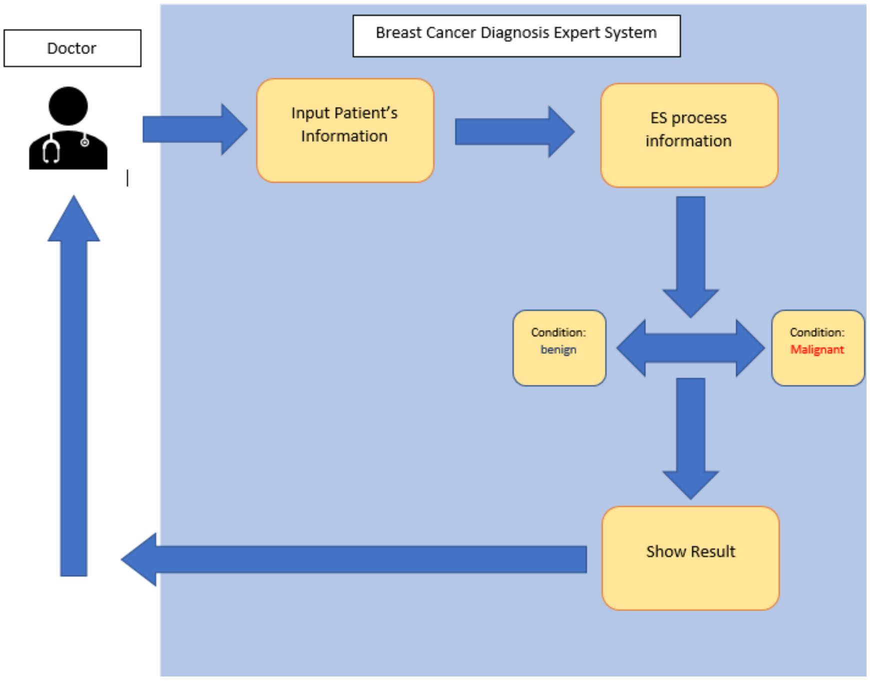


Figure 4.2.3 Use Case Diagram

- Context Diagram

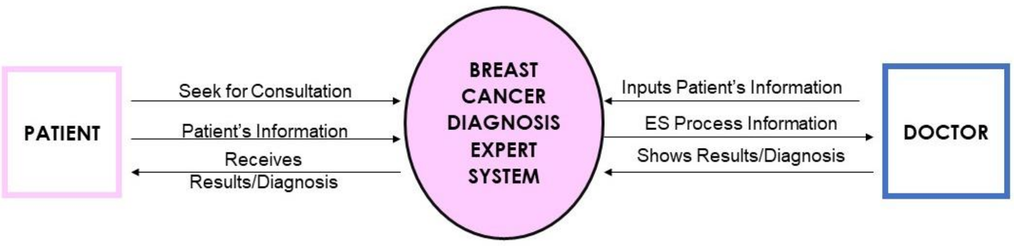


Figure 4.2.4 Context Diagram

4.3 PROJECT IMPLEMENTATION

To implement this simple expert system the IDE used was PyCharm and Jupyter Notebook as well.

4.4 SOFTWARE TESTING

Alpha testing was done with the software, as the developers worked together to came up with the most feasible solutions for the software's needs. Moreover, performance testing was also done, especially with the machine learning model, to ensure its efficacy, quality, and accuracy. User software testing was also conducted to check for any errors or failures in the expert system so that it may be corrected and improved to function properly under its required conditions.

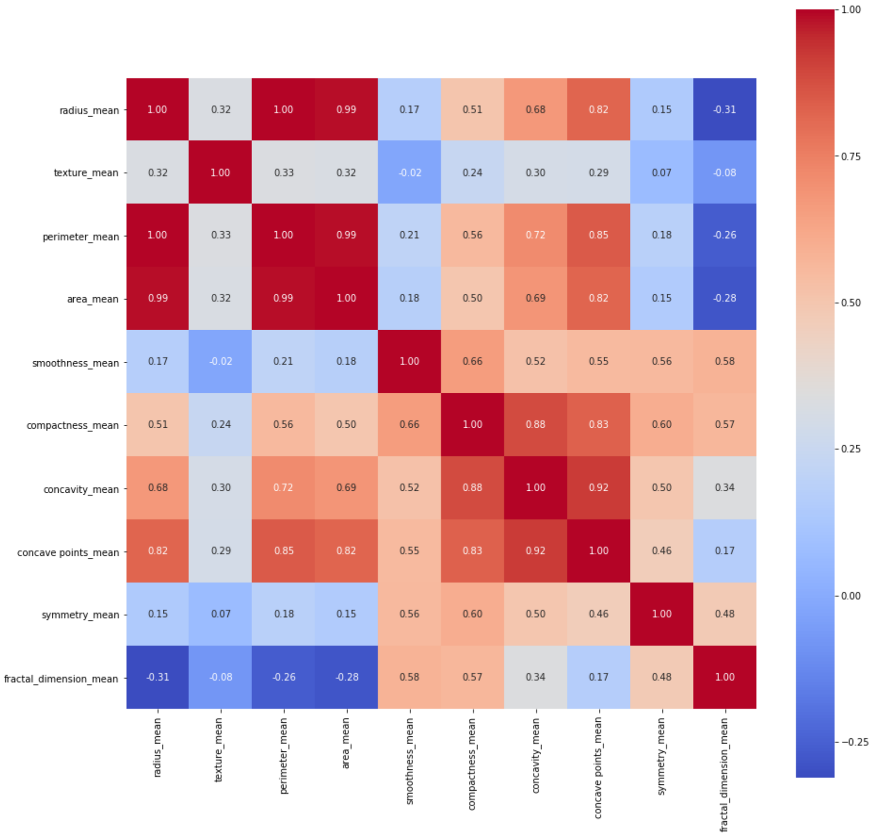
5.0 DISCUSSION OF RESULTS AND FINDINGS

In this chapter, the results of the GUI’s presented as a conclusion of the programming code created using PyCharm and Jupyter Notebook.

5.1 PROJECT IMPLEMENTATION

The following graphs are presented to find and show which variables are correlated to each other. As shown, the values with 1.0 are correlated in the heatmap, data points closer to each other forming a line on our linear regression shows correlation.

- Heatmap for the Mean



	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
radius_mean	1.00	0.32	1.00	0.99	0.17	0.51	0.68	0.82	0.15	-0.31
texture_mean	0.32	1.00	0.33	0.32	-0.02	0.24	0.30	0.29	0.07	-0.08
perimeter_mean	1.00	0.33	1.00	0.99	0.21	0.56	0.72	0.85	0.18	-0.26
area_mean	0.99	0.32	0.99	1.00	0.18	0.50	0.69	0.82	0.15	-0.28
smoothness_mean	0.17	-0.02	0.21	0.18	1.00	0.66	0.52	0.55	0.56	0.58
compactness_mean	0.51	0.24	0.56	0.50	0.66	1.00	0.88	0.83	0.60	0.57
concavity_mean	0.68	0.30	0.72	0.69	0.52	0.88	1.00	0.92	0.50	0.34
concave points_mean	0.82	0.29	0.85	0.82	0.55	0.83	0.92	1.00	0.46	0.17
symmetry_mean	0.15	0.07	0.18	0.15	0.56	0.60	0.50	0.46	1.00	0.48
fractal_dimension_mean	-0.31	-0.08	-0.26	-0.28	0.58	0.57	0.34	0.17	0.48	1.00

Figure 5.1 Mean Heatmap

- Heatmap for the SE

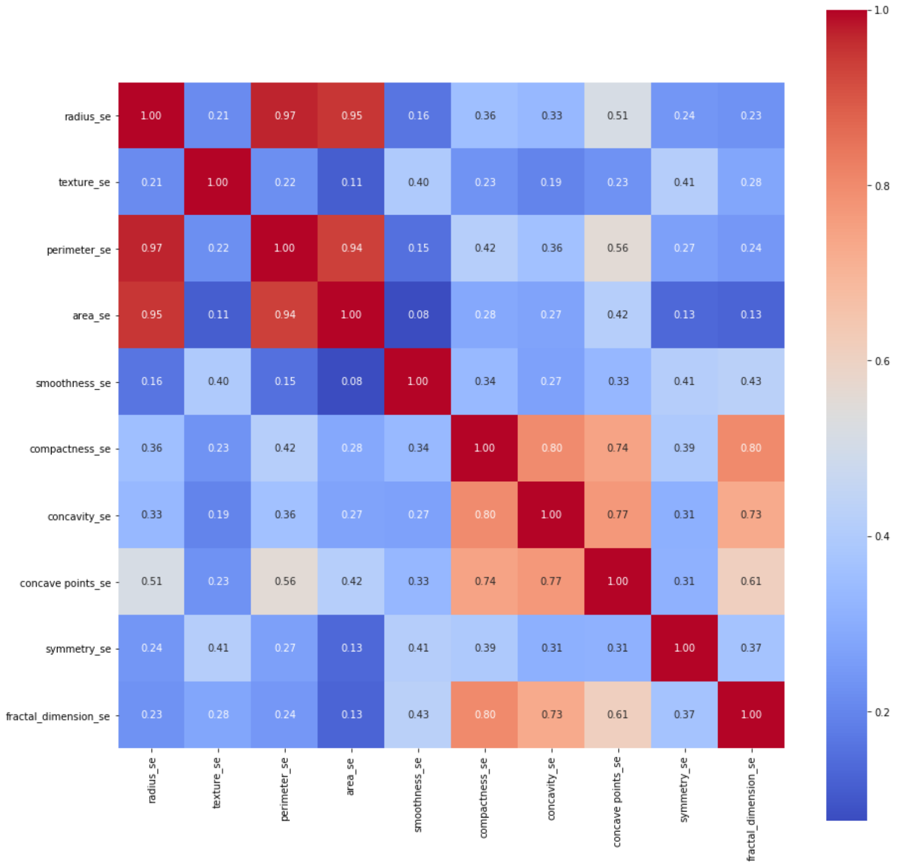


Figure 5.1.2 SE Heatmap

- Heatmap for the Worst

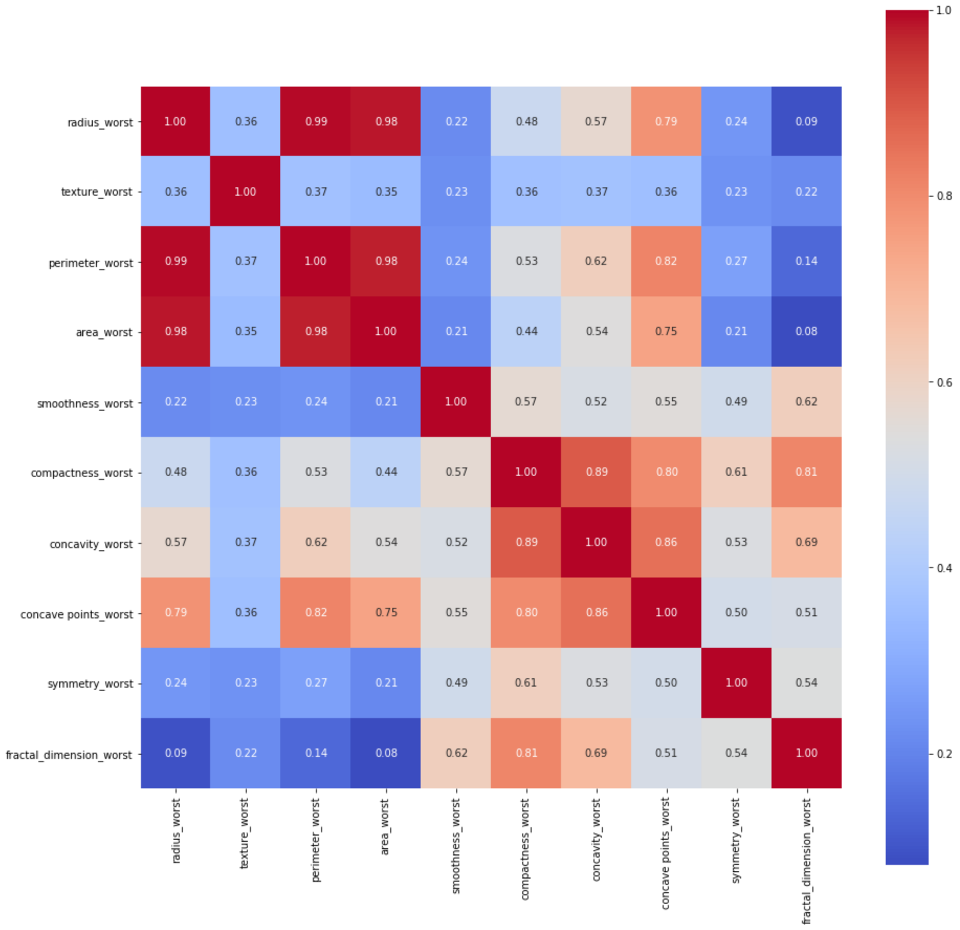


Figure 5.1.3 Worst Heatmap

- Accuracy

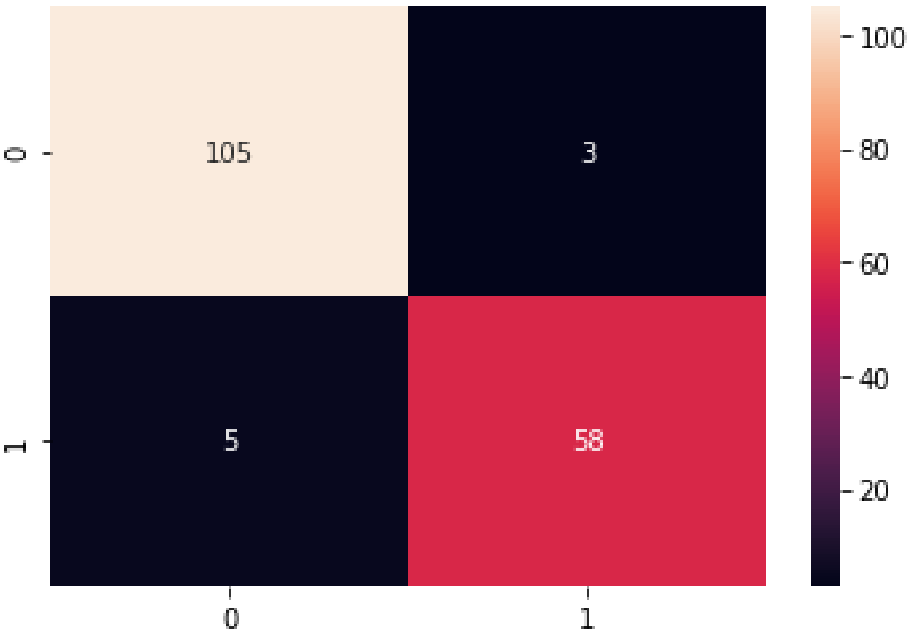


Figure 5.1.4 Accuracy Heatmap

Here, illustrated is the evaluation of our classifier. Our classifier, having 0.9545 precision or correctly classified positive examples divided by the total number of examples that are classified as positive. Our recall, having 0.9722 or the number of correctly classified examples divided by the total number of actual positive examples in the test set. Also, we can see here the F1-score, also known as the harmonic mean of precision and recall, having a 0.9633 value.

- Linear Regression

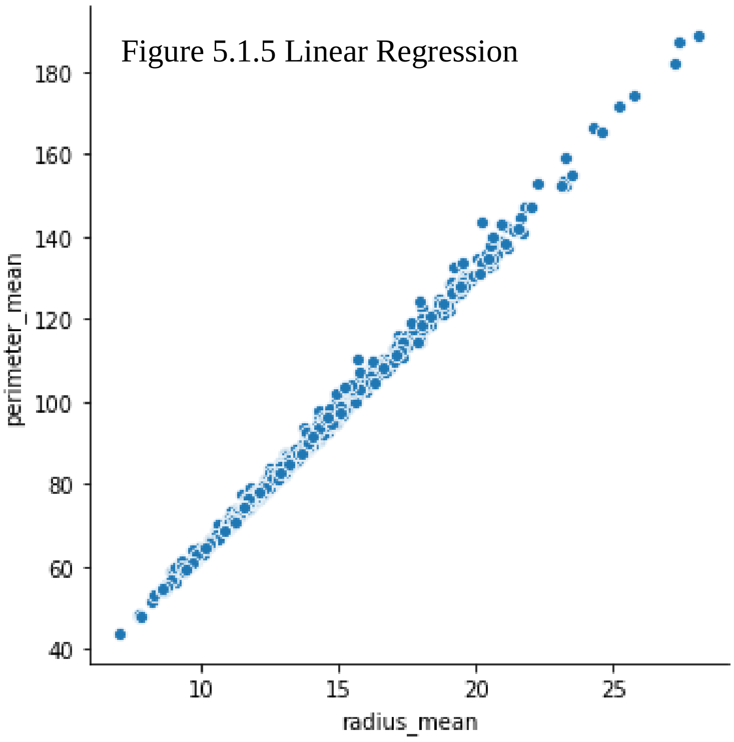


Figure 5.1.5 Linear Regression

Figure 5.1.5 shows the correlation of the Perimeter Mean and Radius Mean. They are the only two variables in the whole dataset that has a perfect degree of correlation, with a correlation value of 1.

- Number of malignant and benign cases in the given dataset

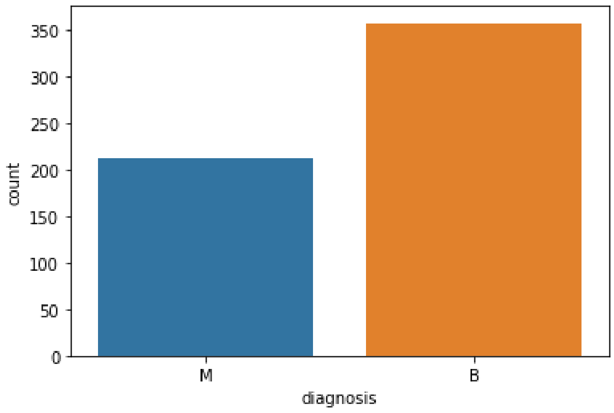
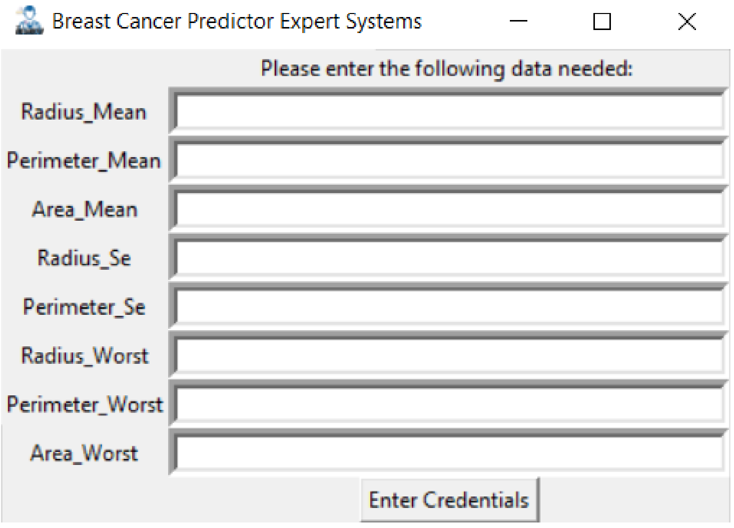
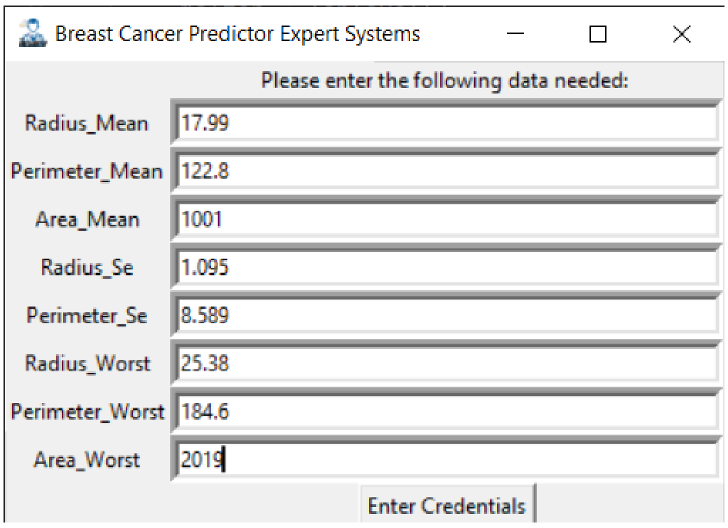


Figure 5.1.6 Number of Malignant and Benign Cases

5.2 SOFTWARE TESTING

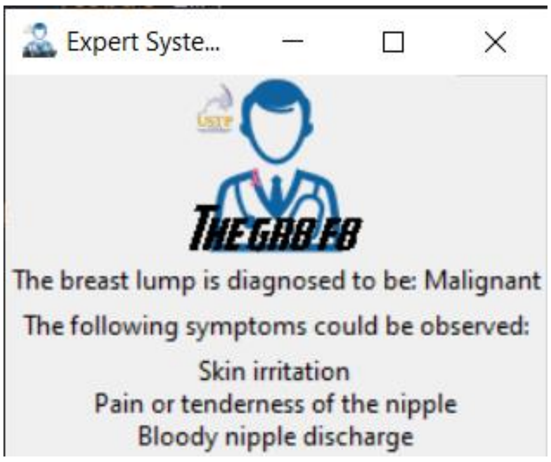
Shown below is the GUI of the breast cancer diagnosis simple expert system, wherein the results are given. If it’s malignant after doing the input, it’ll show with corresponding diagnosis and symptoms, same goes with the benign as well.





Breast Cancer Predictor Expert Systems
Please enter the following data needed:
Enter Credentials

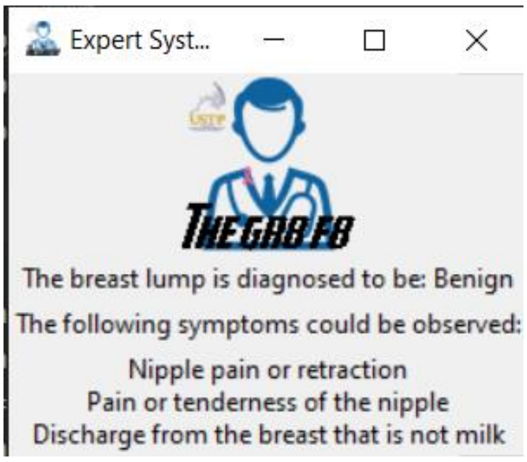
Radius_Mean	17.99
Perimeter_Mean	122.8
Area_Mean	1001
Radius_Se	1.095
Perimeter_Se	8.589
Radius_Worst	25.38
Perimeter_Worst	184.6
Area_Worst	2019



The breast lump is diagnosed to be: Malignant

The following symptoms could be observed:

Skin irritation
Pain or tenderness of the nipple
Bloody nipple discharge



The breast lump is diagnosed to be: Benign

The following symptoms could be observed:

Nipple pain or retraction
Pain or tenderness of the nipple
Discharge from the breast that is not milk

6.0 SUMMARY AND CONCLUSION

6.1 PROJECT DESIGN

Data processes were conducted as followed and showed anticipated outcomes and the developers obtained reliable results. With the aid of the research designs presented, the developers were able to execute the objectives of the expert system.

6.2 PROJECT IMPLEMENTATION

The results were as expected, the ML model gave a high accuracy rate, because also with the help of some data mining and exploratory data analysis. As for the GUI, it easily did its job of giving easy use for users, however, there still a lot of room for improvement for the design.

6.3 SOFTWARE TESTING

With this type of expert system, it's best to be tested by the doctors itself through beta testing and acceptance testing, to identify the flaws of the system and improve it from time to time. Performance testing through validation from different programmers is also a viable step in this.

TDD might be not suitable for this project, as this type of project contains hundreds and even thousands of lines of codes, and if dealt with TDD, time would be the number one concern.

7.0 RECOMMENDATIONS FOR FUTURE PROJECTS BASED ON RESEARCH RESULTS

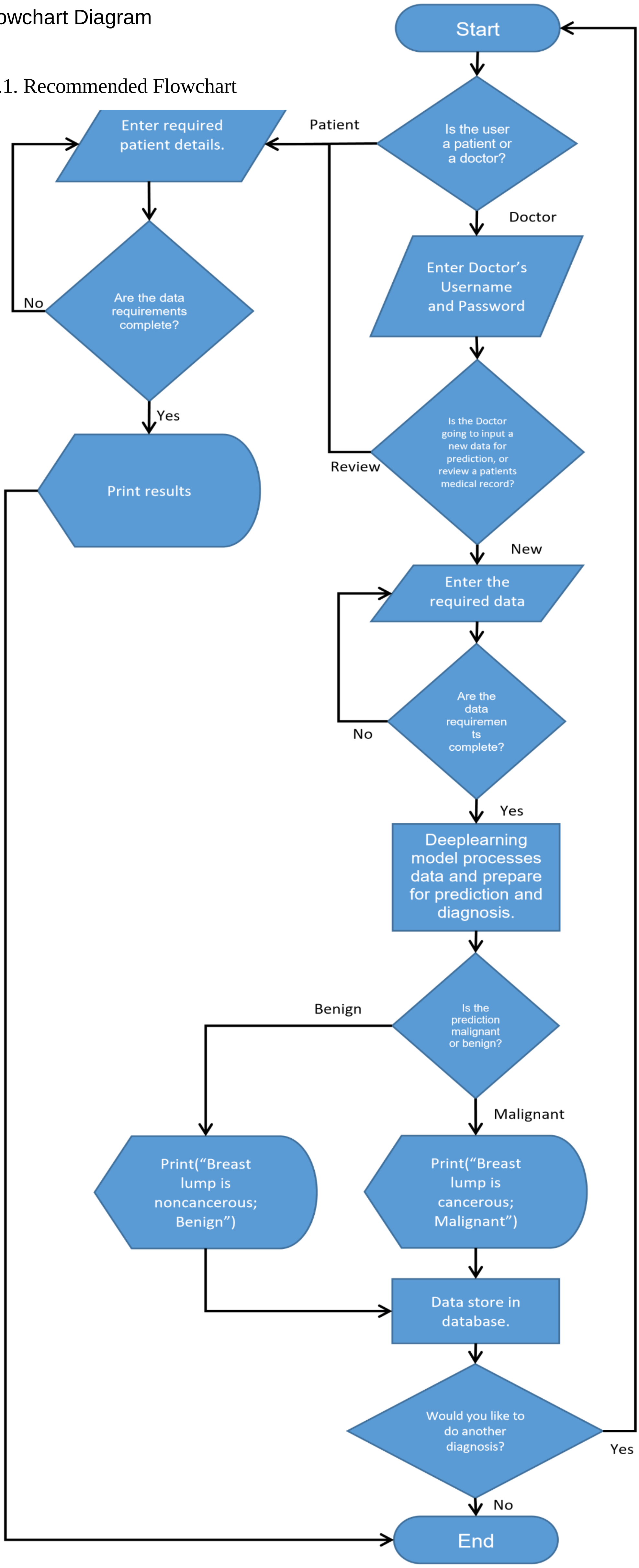
In this chapter suggestions and recommendations are presented to aim for any improvements and developments in the future endeavors to come for the simple expert system.

7.1 PROJECT ARCHITECTURE AND DESIGN

The following are recommended sets of diagrams.

- Flowchart Diagram

Figure 7.1. Recommended Flowchart


Start
Is the user a patient or a doctor?
Patient
Doctor
Enter required patient details.
Are the data requirements complete?
No
Yes
Print results
Enter Doctor's Username and Password
Is the Doctor going to input a new data for prediction, or review a patients medical record?
Review
New
Enter the required data
Are the data requiremen ts complete?
No
Yes
Deeplearning model processes data and prepare for prediction and diagnosis.
Is the prediction malignant or benign?
Benign
Malignant
Print("Breast lump is noncancerous; Benign")
Print("Breast lump is cancerous; Malignant")
Data store in database.
Would you like to do another diagnosis?
Yes
No
End

- Use Case Diagram

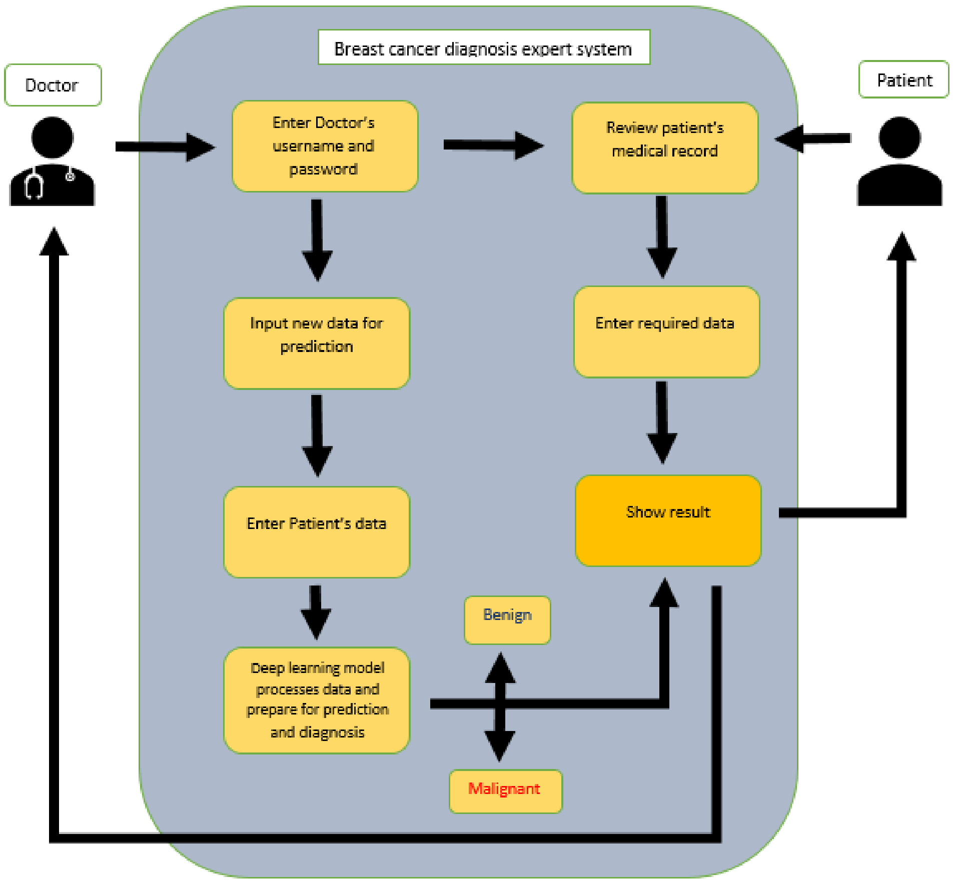

Breast cancer diagnosis expert system
Doctor
Patient
Enter Doctor's username and password
Review patient's medical record
Input new data for prediction
Enter required data
Enter Patient's data
Show result
Benign
Deep learning model processes data and prepare for prediction and diagnosis
Malignant

Figure 7.1.2 Recommended Use Case

- Context Diagram

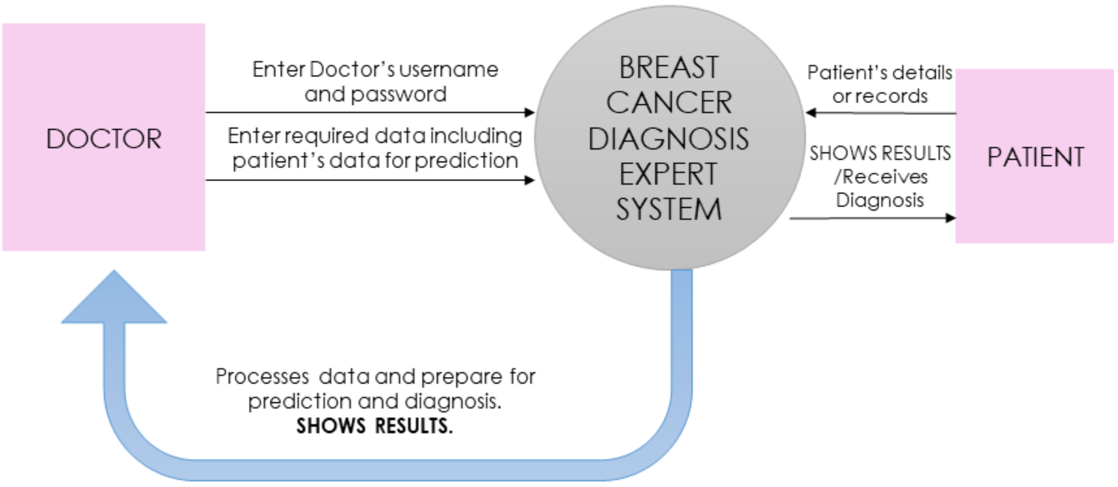

DOCTOR
Enter Doctor's username and password
Enter required data including patient's data for prediction
BREAST CANCER DIAGNOSIS EXPERT SYSTEM
Patient's details or records
SHOWS RESULTS /Receives Diagnosis
PATIENT
Processes data and prepare for prediction and diagnosis.
SHOWS RESULTS.

Figure 7.1.3 Recommended Context

7.2 PROJECT IMPLEMENTATION

The expert system did its job and gave satisfactory results. On the other hand, with this type of research, it's more practical to use a deep learning approach through image segmentation for easier data entry, and more rigid and keen data mining.

7.3 SOFTWARE TESTING

User testing through Alpha Test, Beta Test, and Acceptance test should be mandatory with this type of system, to achieve efficient system function and reliability. Performance testing should also not be taken out of equation, as this plays a vital role in determining the accuracy of the system. TDD might also be considered, if and only if, time is by the researcher’s side**.**

8.0 LITERATURE CITED

- Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis by Wenbin Yue, Zidong Wang, Hongwei Chen, Annette Payne and Xiaohui Liu (https://www.researchgate.net/publication/325064884_Machine_Learning_with_Applications_in_Breast_Cancer_Diagnosis_and_Prognosis)
- Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis Hiba Asri, Hajar Mousannif Hassan, Al Moatassime, Thomas Noel (https://core.ac.uk/download/pdf/82813624.pdf)
- Machine learning applications in cancer prognosis and prediction Konstantina Kouroua, Themis P. Exarchosab, Konstantinos P. Exarchosa, Michalis V. Karamouzisc, Dimitrios I. Fotiadisab (https://www.sciencedirect.com/science/article/pii/S2001037014000464)
- UCI Machine learning Repository on Breast Cancer Wisconsin (Diagnostic) Data Set. (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29?fbclid=IwAR2X6vRKaWgJxN5-pDdNghmRnKxQbRxBIAeIFPjYAsB9K7zTHJsrgE0j1KQ)
- https://www.izenda.com/what-is-healthcare-analytics/
- https://www.managedhealthcareexecutive.com/view/biggest-issues-facing-healthcare-today
- https://reader.elsevier.com/reader/sd/pii/S2001037014000464?token=52D41E75CA9A977226F56E7B4EEB3FEB69FF6F8462DF19934B6C5F86D9134150369D099D7FD256BF33A4E13803981175
- https://www.ncbi.nlm.nih.gov/books/NBK459179/
- https://www.bbc.com/news/health-50857759
- https://www.radiologyinfo.org/en/info.cfm?pg=breastlumps
- https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/limitations-of-mammograms.html
- https://www.ncbi.nlm.nih.gov/books/NBK285/
- https://www.ncbi.nlm.nih.gov/books/NBK285/

- https://www.webmd.com/breast-cancer/breast-self-exam#1
- www.slideshare.net/acijjournal/an-approach-for-breast-cancer-diagnosis-classification-using-neural-network
- www.researchgate.net/publication/329083180_Detection_of_Breast_Cancer_with_Mammography_Effect_of_an_Artificial_Intelligence_Support_System/link/5c50d07f92851c22a39998e1/download
- https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30185-0/fulltext
- https://www.ncbi.nlm.nih.gov/books/NBK225748

9.0 APPENDICES

A. PROBLEMS

- Limited time to meet and discuss about the task online.
- Internet Connection is poor.
- Lack of necessary materials such as laptop.

B. RAW DATA

Dataset was from the UCI Machine learning Repository on Breast Cancer Wisconsin (Diagnostic) Data Set.

CSV File: data.csv