BREAST CANCER DETECTION

A PROJECT REPORT

Submitted by

DEEPASHRI DABHADE- 21BAI10325
SHAUNAK ABHONKAR- 21BAI10348
ANIRUDDHA JOSHI- 21BAI10189
AADITYA MORE- 21BAI10014
MEET JOYSAR- 21BAI10063

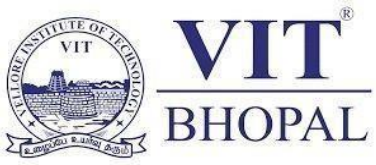
in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

in
COMPUTER SCIENCE AND ENGINEERING

Specialization in

Artificial intelligence and machine learning



SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

VIT BHOPAL UNIVERSITY

KOTHRI KALAN, SEHORE
MADHYA PRADESH - 466114

FEBRUARY 2022

BONAFIDE CERTIFICATE

Certified that this project report titled **"BREAST CANCER DETECTION USING SUPERVISED MACHINE LEARNING"** is the bonafide work of " **DEEPASHRI DABHADE (21BAI10325), SHAUNAK ABHONKAR (21BAI10348), ANIRUDDHA JOSHI (21BAI10189), AADITYA MORE (21BAI10014), MEET JOYSAR (21BAI10063)"** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported at this time does not form part of any other project/research work based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

PROGRAM CHAIR
Dr. S. Suthir, Professor
School of Computer Science and Engineering
VIT BHOPAL UNIVERSITY

PROJECT GUIDE
Dr. Komarasamy G. , Professor
School of Computer Science and Engineering
VIT BHOPAL UNIVERSITY

The Project Exhibition II Examination is held on 26th December 2022 to 16th February 2023.

ACKNOWLEDGEMENT

First and foremost I would like to thank the Lord Almighty for His presence and immense blessings throughout the project work.

We wish to express our heartfelt gratitude to Dr S. Suthir, Head of the Department, School of Computer Science and Engineering for much of his valuable support encouragement in carrying out this work.

We would like to thank our internal guide Mr. Komarasamy G. ,for continually guiding and actively participating inour project, giving valuable suggestions to complete the project work.

We would like to thank all the technical and teaching staff of the School of Aeronautical Science, who extended directly or indirectly all support.

Last, but not least, We are deeply indebted to our parents who have been the greatest support while we worked day and night for the project to make it a success.

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Currently, Breast cancer (BC) has a relatively high mortality rate and is the second most common kind of cancer in women that causes death. Its impact will be lessened by early discovery. Early BC diagnosis may encourage patients to undergo expedited surgical treatment, which will greatly enhance the prognosis and likelihood of recovery. Hence, creating a system that enables the healthcare industry to quickly and accurately diagnose breast cancer is crucial. Machine learning (ML) is commonly used in the categorization of breast cancer (BC) pattern due to its advantages in modelling a crucial feature detection from complex BC datasets. For the automatic diagnosis and prognosis of BC, we suggest an ensemble of classifier-based systems in this research. Breast cancer symptoms include breast tumours, bloody nipple discharge, and changes to the texture of the breast or nipple. Breast cancer treatments include surgery, hormone replacement therapy, chemotherapy, and radiation, depending on the stage. The objective of this project is to develop a machine learning model to classify breast cancer as benign or malignant. The methodology used for the model includes random forests, decision trees, and logistic regression. SVM (Support Vector Machine) provided us with the highest accuracy, or 98.2%.

TABLE OF CONTENTS

INTRODUCTION

1.1 INTRODUCTION

WHAT IS BREAST CANCER?

- Breast cancer develops in the tissue of the breast.
- Indications of breast cancer include breast lumps, bloody nipple discharge, and changes in nipple or breast texture.
- Treatment for breast cancer is determined by its stage and may include chemotherapy, radiation, hormone replacement treatment, and surgery.

WHAT IS CANCER DETECTION?

- Cancer cannot be accurately diagnosed by a single test. A thorough history and physical examination, as well as diagnostic tests, are typically necessary for a patient's full evaluation.
- To assess whether a person has cancer or whether another disorder (such an infection) is imitating the symptoms of cancer, numerous tests are required.
- Effective diagnostic testing is used to confirm or rule out the presence of disease, track the progression of the condition, and organize and assess treatment. When a patient's condition changes, a poor-quality sample was obtained, or an aberrant test result needs to be validated, it may be essential to repeat the test.
- Imaging, laboratory tests (including tests for tumor markers), tumor biopsy, endoscopic examination, surgery, or genetic testing are all examples of diagnostic techniques for cancer.

MOTIVATION OF WORK

We found that many of the Cancer Detecting procedures were extremely time consuming and Cumbersome. We are aiming to create a system that will be easier to operate and understand and will the give the results with much more accuracy than the other systems.

PROBLEM STATEMENT

One of the leading causes of cancer-related deaths globally is breast cancer. Early diagnosis greatly increases the likelihood of receiving the proper care and surviving, however this procedure is time-consuming and frequently causes conflict between pathologists. Systems for computer-aided diagnostics had the potential to increase diagnostic precision. Yet, early detection and prevention can greatly lower the risk of death. Finding breast cancer as soon as possible is crucial.

With the aid of machine learning models, we will identify and detect the Breast Cancer of a listed patient in this project. We will examine data that has been obtained through Fine Needle Aspiration biopsy procedure from a pathological lab. We will work upon the dataset with algorithms.

PROPOSED MODEL

By using biopsy data for training and testing, we have developed a machine learningsystem. Our model recognizes the type of Breast Cancer and alerts the appropriate authorities. To complete the job, we have employed a variety of strategies. We have performed data processing and data analysis on the data. The model uses Random Forest, Decision Trees, Support Vector Machine and Logistic Regression. We attempt to find out the type of Breast Cancer and thus create an easier way of detection for the Cancer.

LITERATURE SURVEY

INTRODUCTION

Several automated systems that use different algorithms have emerged in recent years to categorise breast cancer. The extraction of distinguishing characteristics is necessary before classification in order to classify breast cancer.

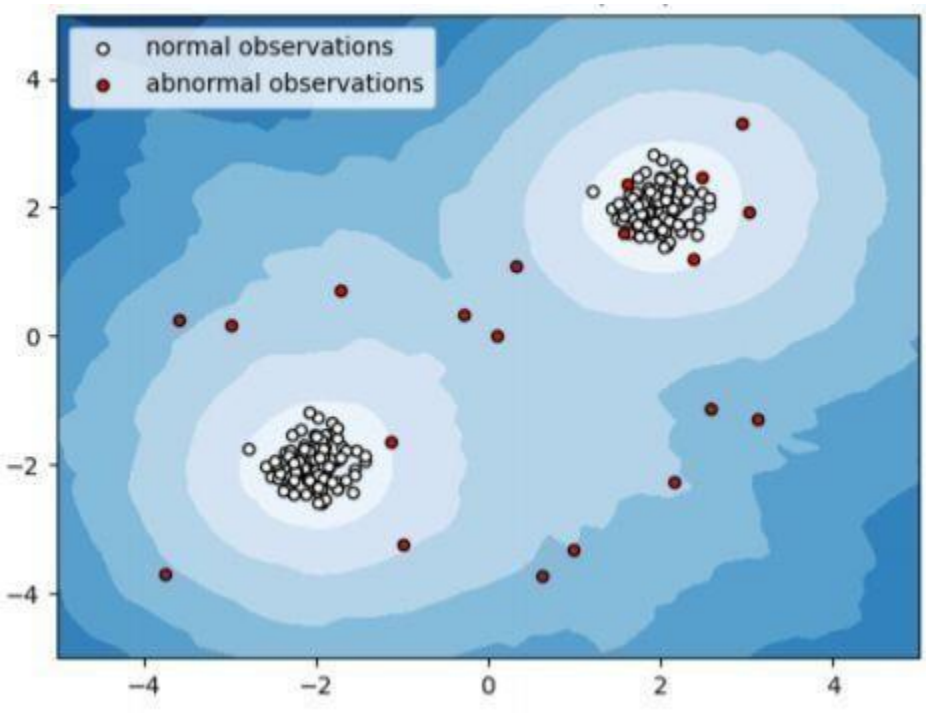
Whitaker et al. suggested a two-stage patch classification approach for mammography employing the texture descriptors "Pass Band Discrete Cosine Transform (PB-DCT)" and "Histogram of Oriented Texture (HOT)". In the beginning, mammogram patches are classified as normal or abnormal.

Using known samples from a training dataset, nearest neighbor algorithms classify the data by locating its closest neighbors in a multidimensional feature space. With increased dimension ratios for the closest neighbors, forecasting accuracy rises. Because the results of this technique depend on how the distance between the data is determined, both the Manhattan distance and the Euclidean distance were examined at this time.

EXISTING ALGORITHMS

A. LOCAL OUTLIER FACTOR

It is an Unsupervised Outlier Detection algorithm. 'Local Outlier Factor' refers to the anomaly score of each sample. It measures the local deviation of the sample data with respect to its neighbours. More precisely, locality is given by k-nearest neighbours, whose distance is used to estimate the local data.



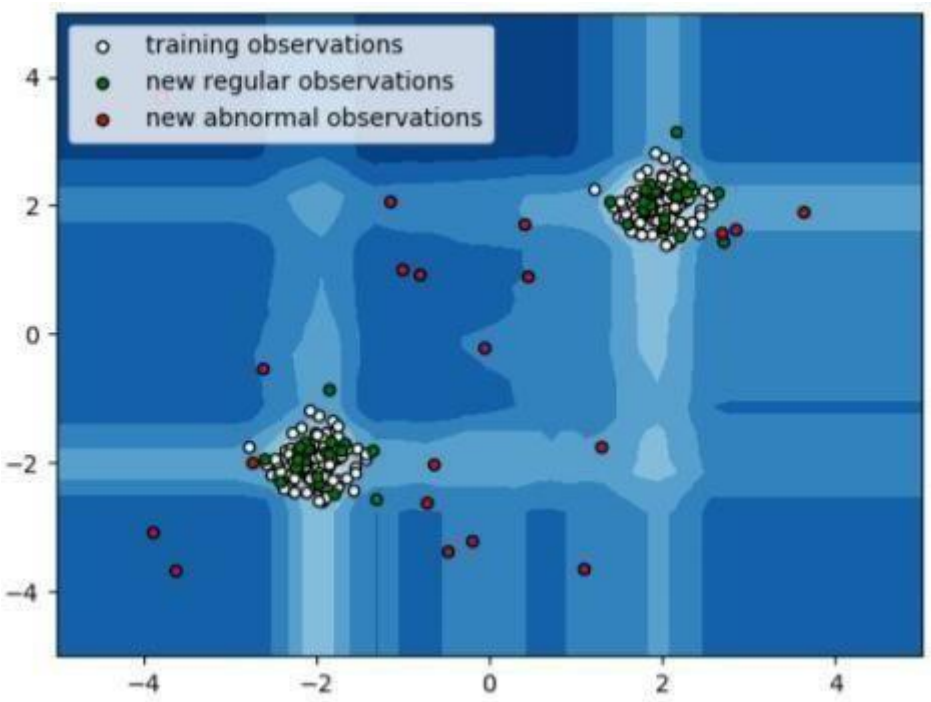
B. ISOLATION FOREST

The Isolation Forest ‘isolates’ observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.

Since recursive partitioning can be represented by a tree structure, the number of splittings required to isolate a sample is equivalent to the path length from the rootnode to the terminating node.

This path length, averaged over a forest of such random trees, is a measure of normality and our decision function.

Random partitioning produces noticeably shorter paths for anomalies. Hence, when a forest of random trees collectively produce shorter path lengths for particular samples, they are highly likely to be anomalies.



RESEARCH ISSUES/OBSERVATIONS FROM LITERATURE SURVEY

Of the many major issues facing the graduate student, a primary one is the identification of a research problem. Problems may arise from real-world settings or be generated from theoretical frameworks. The source of research problems will vary according to the experience of the person contemplating an investigation, but it is generally agreed that the process begins with a question or need.

Title of the paper	Journal Name, Publisher Name, Year of Publication and Volume & Issue Number (only SCI)	Author Name	Problem addressed/ Problem Statement	Methods/Technologies used	Author Contribution	Shortcoming/ Assumption Made
An Automatic Detection of Breast Cancer Diagnosis and Prognosis Based on Machine Learning Using Ensemble of Classifiers	IEEE Access (Vol 10), IEEE, 12 May 2022, *Funding Agency:* Technology Development Program of MSS, National Research Foundation of Korea (NRF)	Usman Naseem, Junaid Rashid, Liaqat Ali, Jungeun Kim, Qazi Emad Ul Haq, Mazhar Javed Awan, Muhammad Imran	Providing the ML model such that it would have great accuracy with a new approach rather than following the other state-of-the-art methods	#. A classification framework was created using an ensemble of the machine learning classifiers SVM, LR, NB, and DT.	Presented an ensemble of machine learning-based methods for breast cancer diagnosis and prognosis.	The performance of the classifiers could vary depending on the specific hyperparameter values chosen. It may be beneficial to perform a hyperparameter search in order to find the optimal values for each classifier.

DATASET

We have derived our dataset from GitHub. It consists of 33 columns and 569 rows. We have over 212 Breast Cancers detected. The '1' in Class feature indicates a Malignant Tumor and '0' indicates Benign Tumor.

METHODOLOGIES

Importing Dependencies:

In this step, we import the different modules and libraries that we will require to continue working on the project.

Exploratory Data Analysis:

In this method, several logical and machine learning approaches are used to analyze and work with the data.

Data division into train and test data:

Two categories of datasets are created for machine learning. A portion of our actual dataset that is utilized to train a machine learning model is the first subset, also known as the training data. In this way, it trains our model.

After building your machine learning model, you need to test it using unknown data (using your training data). This data, also known as testing data, can be used to evaluate the efficiency and growth of the training of your algorithms and to adjust or optimize them for better results.

Building a Model:

In this step, we build a model by generalizing the knowledge it has learned from training data and utilizing it to generate predictions and accomplish its objective.

Calculating the Accuracy of the Models and Comparing them:

Here, we calculate the accuracies of the various models implemented in the Machine Learning Coding and then we compare and conclude on the best Model.

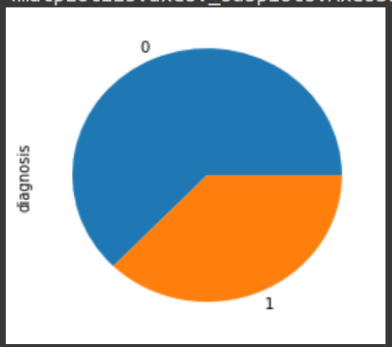
EXPERIMENTAL SETUP

- Initially, we gathered data from Kaggle and then worked upon it.
- Then, as advised by our mentor, we studied around 10-15 research papers regarding thesimilar topic.
- We then enhanced our knowledge on machine learning by completing a few courses online and understanding techniques and models.
- We understood Logistic Regression, Decision Trees, Random Forest, and SVM and further began working on it.
- We then tried to perform analysis on the data and implemented it in the code.
- We finally completed the coding part with our mentor's guidance and teachings.
- We completed the presentation and report based on the guidelines provided by the University.

OUTPUT SCREEN

```
# Checking the class distribution of the target variable in percentage
print((df.groupby('diagnosis')['diagnosis'].count()/df['diagnosis'].count()) *100)
((df.groupby('diagnosis')['diagnosis'].count()/df['diagnosis'].count()) *100).plot.pie()
```

```
diagnosis
0    62.741652
1    37.258348
Name: diagnosis, dtype: float64
<matplotlib.axes._subplots.AxesSubplot at 0x7f5dd1375a30>
```



This graph shows us the representation of Benign and Malignant Tumors

```
def models(X_train,Y_train):
        #logistic regression
        from sklearn.linear_model import LogisticRegression
        log=LogisticRegression(random_state=0)
        log.fit(X_train,Y_train)


        #Decision Tree
        from sklearn.tree import DecisionTreeClassifier
        tree=DecisionTreeClassifier(random_state=0,criterion="entropy")
        tree.fit(X_train,Y_train)
```

Here we begin the coding part for implementing Logistic Regression and Decision Tree.

SYSTEM ANALYSIS

INTRODUCTION

This chapter gives the information regarding analysis done for the proposed system. System Analysis is done to capture the requirement of the user of the proposed system. It also provides the information regarding the existing system and also the need for the proposed system. The key features of the proposed system and the requirement specifications of the proposed system are discussed below.

EXISTING SYSTEM

The Traditional detection method mainly depends on the database system and the education of customers, which usually are delayed, inaccurate and not in-time. After that methods based on discriminant analysis and regression analysis are widely used which can detect Cancer. For a large amount of data it is not efficient.

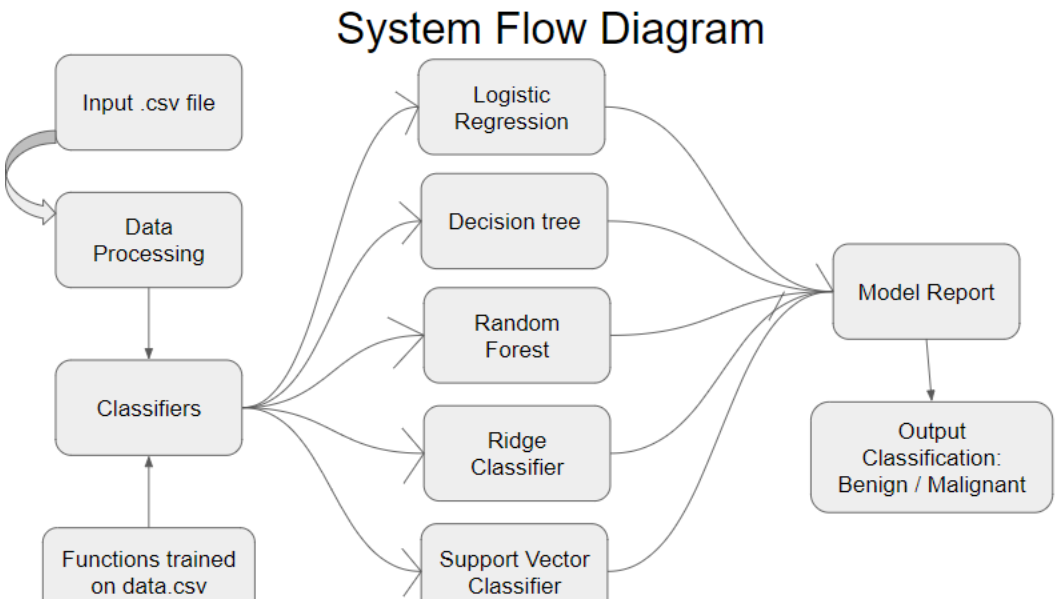
PROPOSED SYSTEM

The proposed system overcomes the above mentioned issue in an efficient way. Using supervised machine learning algorithms the cancer is detected and the false alert is minimized and it produces an optimized result. The cancer is detected based on the cell's behavior. A new classification problem which has a variable misclassification cost is introduced. Here the SVM algorithm is made where a set of interval valued parameters are optimized

SYSTEM DESIGN AND IMPLEMENTATION

ARCHITECTURAL DESIGN:

Defining the requirements and setting the high level of the system are both concerns when describing the general aspects of the software. The different web pages and their connections are recognized and designed during architectural design. The main parts of the software are recognized, broken down into processing modules and conceptual data structures, and their linkages are shown. The suggested system includes the following modules.



The above architecture describes the work structure of the system.

The data is prioritized by the filter and priority module before being sent to the genetic algorithm, which carries out its operations and produces the output.

DETAILED SYSTEM DESIGN :

The various modules are covered in full in the detailed design, with the necessary diagrams and notations. The use case diagram is intended to show the proposed system's operational logic. The purpose of the sequence diagram is to explain how the client and server work together to process content. The activity diagram is used to describe the proposed system's flow. After analyzing the keywords and the current URL link, we are aware of the beginning and ending points of the program. This will make it easier for the programmers to implement the module's internal logic in the specified specification.

Top-down design methodology is used in this stage of the design process. The primary modules are first determined. They are then split up into smaller modules, with the lowest level of each module addressing a specific system function. The details of each module design are provided. This chapter explains how the input module is created to meet the needs of the user. What technologies are used to collect inputs and send them to the server is explained in the full input design.

The user has several good screen-interaction options thanks to output design. the data made available to users via an information system. To ensure that the information system is used and accepted, useful output is crucial. Users frequently assess a system's worth depending on its results. Only close interaction with users will result in productive productivity. The output is created in a beautiful and efficient manner so users may access them when they have an issue.

IMPLEMENTATION

The suggested machine-learning algorithms could predict breast cancer since timely therapeutic interventions could assist slow the progression of the disease and lower mortality rates through early diagnosis of this condition. The performance of modelling could be enhanced by using various machine learning techniques, having access to larger datasets from several institutions (multi-center study), and taking into account important attributes from a number of pertinent data sources.

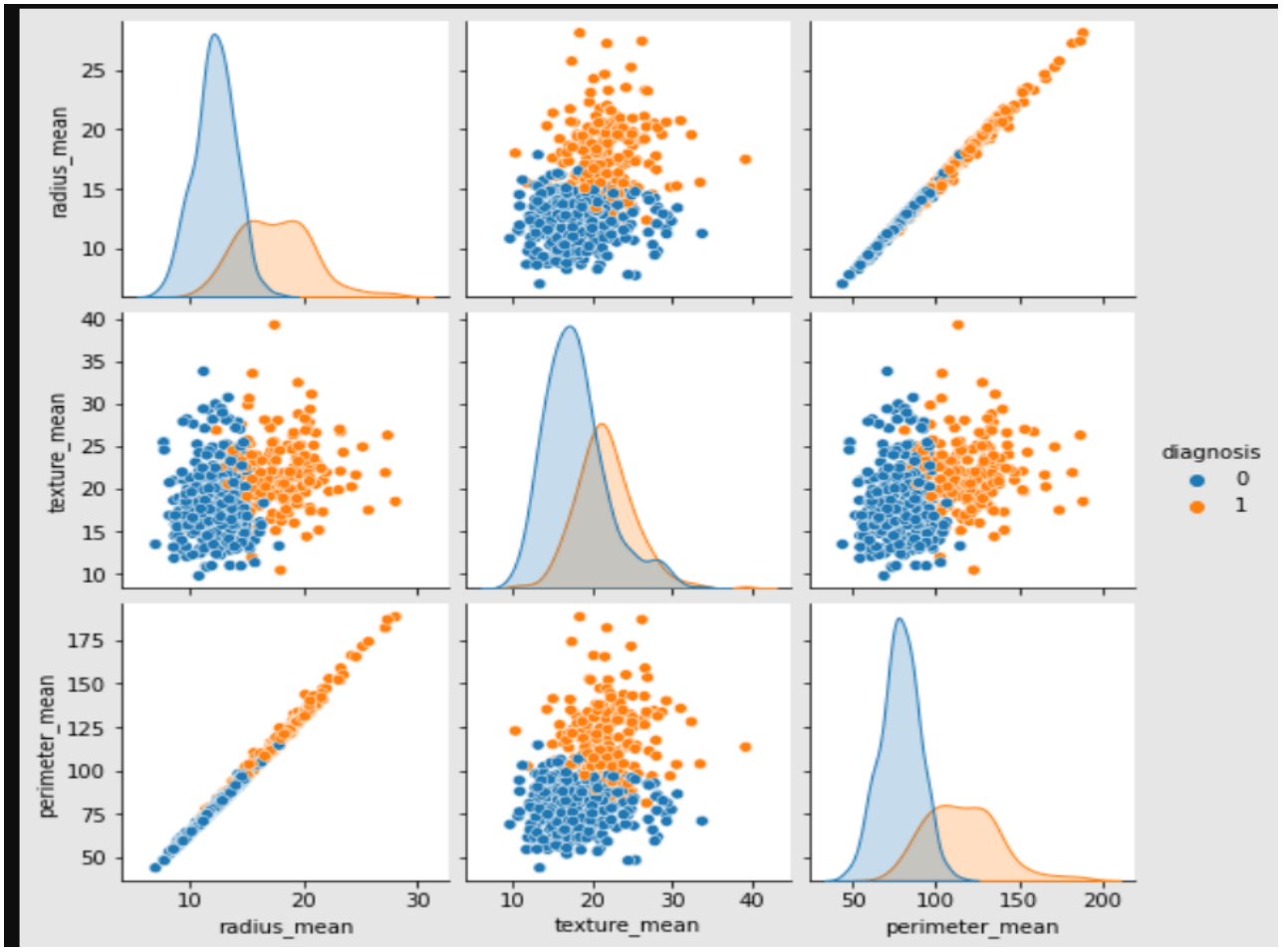
HARDWARE AND SOFTWARE REQUIREMENTS:

S.No	Hardware used	Justification
1	Laptop (Used : Ryzen 7 5800H, RAM : 8GB, SSD : 256GB)	• Development and Testing of the model • portability

S.No	Software used	Justification
1	Python 3.6.9 from Google Collaboratory	• Google colab pre-includes the Scipy, Numpy, Pandas, Sklearn, Matplotlib and various Machine Learning libraries. • It also gives us a free virtual machine on cloud with capabilities of training ML models

CODING AND TESTING

Testing the DataSet(Graphs)



Implementation of Algorithms using Python

```python
# models/ Algorithms

def models(X_train,Y_train):
        #logistic regression
        from sklearn.linear_model import LogisticRegression
        log=LogisticRegression(random_state=0)
        log.fit(X_train,Y_train)


        #Decision Tree
        from sklearn.tree import DecisionTreeClassifier
        tree=DecisionTreeClassifier(random_state=0,criterion="entropy")
        tree.fit(X_train,Y_train)

        #Random Forest
        from sklearn.ensemble import RandomForestClassifier
        forest=RandomForestClassifier(random_state=0,criterion="entropy",n_estimators=10)
        forest.fit(X_train,Y_train)
```

```
Model 1 -> Logistic regression :
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        67
           1       1.00      0.91      0.96        47

    accuracy                           0.96       114
   macro avg       0.97      0.96      0.96       114
weighted avg       0.97      0.96      0.96       114

Accuracy :  0.9649122807017544

Model 2 -> Decision tree :
              precision    recall  f1-score   support

           0       0.95      0.93      0.94        67
           1       0.90      0.94      0.92        47

    accuracy                           0.93       114
   macro avg       0.93      0.93      0.93       114
weighted avg       0.93      0.93      0.93       114

Accuracy :  0.9298245614035088
```

Future Enhancements And Conclusion

Introduction

The findings obtained here may not be generalized to the global cancer detection problem. As future work, some effective algorithms which can perform well for the classification problem with variable misclassification costs could be developed.

Limitation/Constraints of the System

- Hardware Limitations: There are no hardware limitations.
- Interfaces to other Applications: There shall be no interfaces.
- Parallel Operations: There are parallel operations.
- Audit Functions: There shall be no audit functions.
- Control Functions: There shall be no control functions.

Conclusion

In the oversample cases, of all the models we build found that the SVM model gave us the best accuracy and on oversampled data. We have received the below metrices :
SVM Model: 97.36 % accuracy is by far the best one that we have received.
However, of all the models we created we found the Support Vector Classifier gave us the best result.

Result

```
✓ [28] Accuracy :  0.956140350877193
0s

        Model 5 -> Support Vector Classifier :
                      precision    recall  f1-score   support

                   0       0.97      0.99      0.98        67
                   1       0.98      0.96      0.97        47

            accuracy                           0.97       114
           macro avg       0.97      0.97      0.97       114
        weighted avg       0.97      0.97      0.97       114

        Accuracy :  0.9736842105263158
```

REFERENCES

1. Yanan Shao, Hoda S. Hashemi, Paula Gordon, Linda Warren, Jane Wang, Fellow, IEEE, Robert Rohling, Fellow, IEEE, and Septimiu Salcudean, “Breast Cancer Detection using Multimodal Time Series Features from Ultrasound Shear Wave Absolute Vibro-Elastography”, Published in: IEEE Journal of Biomedical and Health Informatics Volume: 26, Issue: 2, February 2022, DOI: 10.1109.
2. Ioannis Iliopoulos, Simona Di Meo, Marco Pasian, Maxim Zhadobov, Philippe Pouliguen, Patrick Potier, Luca Perregrini, Ronan Sauleau, and Mauro Ettorre, “Enhancement of Penetration of Millimetre Waves by Field Focusing Applied to Breast Cancer Detection”, Published in: IEEE Transactions on Biomedical Engineering (Volume: 68, Issue: 3, March 2021), ISSN: 1245-41508,DOI: 23.1765.
3. Nan Wu; Jason Phang; Jungkyu Park; Yiqiu Shen; Zhe Huang; Masha Zorin, “Deep Neural Networks Improve Radiologists’ Performance in Breast Cancer Screening”, N. Wu et al., "Deep Neural Networks Improve Radiologists’ Performance in Breast Cancer Screening," in IEEE Transactions on Medical Imaging, vol. 39, no. 4, pp. 1184-1194, April 2020, ISSN: 2169-3536, DOI: 10.1109
4. P. Esther Jebarani; N. Umadevi; Hien Dang; Marc Pomplun, “A Novel Hybrid K-Means and GMM Machine Learning Model for Breast Cancer Detection”, P. E. Jebarani, N. Umadevi, H. Dang and M. Pomplun, "A Novel Hybrid K-Means and GMM Machine Learning Model for Breast Cancer Detection," in IEEE Access, vol. 9, pp. 146153-146162, 2021
5. Meriem Sebai; Tianjiang Wang; Saad Ali Al-Fadhli, “Part Mitosis: A Partially Supervised Deep Learning Framework for Mitosis Detection in Breast Cancer Histopathology Images”, IEEE Access, vol. 8, pp. 45133-45147, 2020, ISSN: 2169-3536,DOI: 10.1109
6. Usman Naseem, Junaid Rashid, Liaqat Ali, Jungeun Kim, Qazi Emad Ul Haq, Mazhar Javed Awan,Muhammad Imran, “An Automatic Detection of Breast Cancer Diagnosis and Prognosis Based on Machine Learning Using Ensemble of Classifiers”, IEEE Access (Vol 10), IEEE, 12 May 2022, DOI: 10.1109/ACCESS.2022.3174599, ISSN: 2169-3536
7. Khairul Munadi, Biswajeet Pradhan,Maimun Syukri, Roslidar Roslidar, Aulia Rahman, Rusdha Muharar, Muhammad Rizky Syahputra, Fitri Arnia, “A Review on Recent Progress in Thermal Imaging and Deep Learning Approaches for Breast Cancer Detection”, IEEE Access (Volume: 8), IEEE, 22 June 2020, ISSN: 2169-3536, DOI: 10.1109
8. Yi Wang, Na Wang, Min Xu, Junxiong Yu, Chenchen Qin, Xiao Luo, Xin Yang, Tianfu Wang, Anhua Li, and Dong N, “Deeply-Supervised Networks With Threshold Loss for Cancer Detection in Automated Breast Ultrasound”, doi:10.1109

9. Jing Zheng, Denan Lin, Zhongjun Gao, Shuang Wang, Mingjie He, And Jipeng Fan, “Deep Learning Assisted Efficient AdaBoost Algorithm for Breast Cancer Detection and Early Diagnosis”, Published in: IEEE Access (Volume: 8) , 8 May 2020, ISSN: 2169-3536, DOI: 10.1109
10. Gege Ma and Manuchehr Soleimani, “Spectral Capacitively Coupled Electrical Resistivity Tomography for Breast Cancer Detection”, Published in: IEEE Access (Volume: 8), 11 March 2020, Electronic ISSN: 2169-3536,DOI: 10.1109