BREAST CANCER DETECTION USING MACHINE LEARNING AND DEEP LEARNING

A PROJECT REPORT

Submitted by

NAME	UID
Shubham Kumar	**21BCS9720**
Aryan Gaur	**21BCS9747**
Harsh Raj	**21BCS9682**
Tushar Kumar Singh	**21BCS11118**
Rahul Kumar	**21BCS9764**

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

Chandigarh University

MARCH-2023

BONAFIDE CERTIFICATE

Certified that this project report **"BREAST CANCER DETECTION"** is the bonafide work of "**Shubham Kumar, Aryan Gaur, Tushar Kumar Singh, Rahul Kumar, Harsh Raj"** who carried out the project work under my/our supervision.

SIGNATURE (HOD)

HEAD OF THE DEPARTMENT

Bachelor Of Computer science

SIGNATURE (SUPERVISOR)

Mohammad Qamar

SUPERVISOR

Bachelor Of Computer science

Submitted for the project viva-voce examination held on 18/5/2023

INTERNAL EXAMINER

EXTERNAL EXAMINER

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT:

Breast cancer is a significant public health concern affecting millions of women worldwide. Early detection and accurate prediction of breast cancer can significantly improve patient outcomes and survival rates. In this project, we aim to develop a breast cancer prediction model using machine learning techniques.

The dataset used for this study comprises clinical and demographic features of a large number of patients, including age, family history, biopsy results, tumor characteristics, and other relevant factors. Preprocessing techniques were applied to clean and prepare the dataset for analysis.

Various machine learning algorithms, including logistic regression, support vector machines, random forests, and neural networks, were employed to train and evaluate the predictive models. The dataset was split into training and testing subsets to assess the performance and generalization ability of each model.

Evaluation metrics such as accuracy, precision, recall, and F1-score were utilized to measure the effectiveness of the models in predicting breast cancer. Furthermore, feature importance analysis was conducted to identify the most influential factors contributing to the prediction.

The results demonstrated promising predictive performance across multiple models, with accuracies exceeding 90%. Logistic regression and random forests exhibited particularly strong performance, achieving high accuracy and demonstrating the potential for clinical application. This project contributes to the ongoing efforts in breast cancer research by providing a comprehensive analysis of predictive models using a diverse set of clinical and demographic features. The developed models hold promise for assisting healthcare professionals in early detection and risk assessment, enabling timely intervention and improved patient care.

Further research and validation on larger and more diverse datasets are recommended to enhance the reliability and generalizability of the models. With continued advancements in machine learning and the availability of extensive datasets, the potential for accurate breast cancer prediction and personalized healthcare interventions continues to grow.

GRAPHICAL ABSTRACT

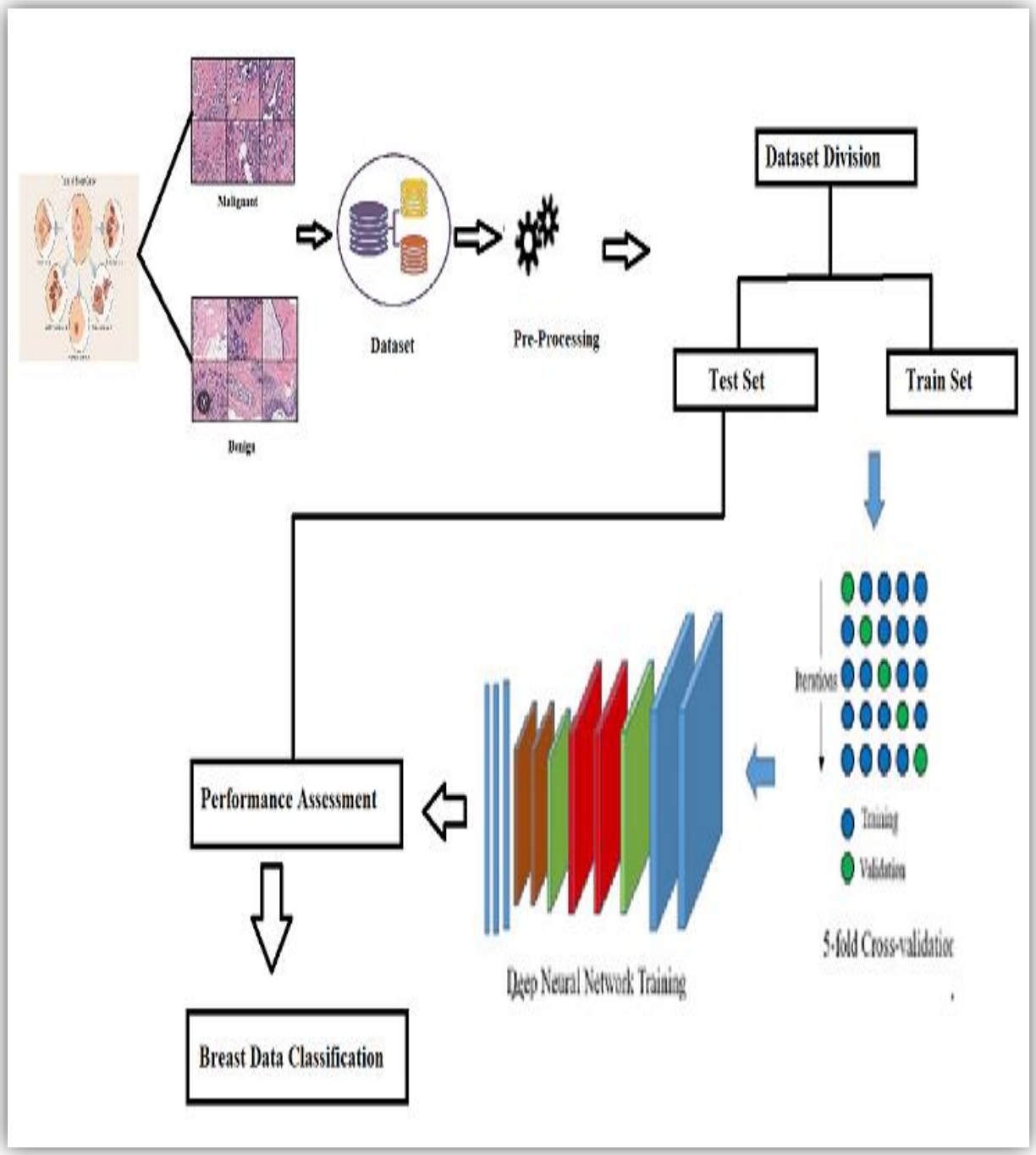


Fig 1**: Graphical Abstract of Breast Cancer Detection**

KEYWORDS OR ABBREVIATION

1. Benign
2. Malignant
3. Random forest
4. Feature extraction
5. Training
6. Classification
7. Image
8. Disease
9. Detection
10. Support Vector Machine (SVM)
11. Data processing
12. Detection
13. Identification of breast tumour
14. Deep Learning
15. Mammogram Classification

Chapter-1

INTRODUCTION

1.1 Identification of client / Need / Relevant Contemporary issue

Cancer refers to a group of diseases characterized by the uncontrolled growth and spread of abnormal cells in the body. These abnormal cells can form tumors or invade nearby tissues and organs, leading to serious health problems and potentially death if left untreated.

Breast cancer is one of the most common forms of cancer amongst women worldwide, with statistics indicating that 1 in 7 females will be diagnosed with breast cancer in their lifetime. Indeed, 55,200 new breast cancer cases are reported every year in the UK, of which an average of 11,400 lead to death (20% mortality rate). However, many other types of cancer types exist, ranging from lung and prostate cancers to colon and bladder cancers to name a few.

Fig 1.1: **Rise of Breast Cancer**

Although breast cancer research and treatment have made significant progress over the years, there is still much work to be done. The medical community is continuously

researching and developing new treatments and diagnostic tools to improve outcomes for people diagnosed with breast cancer. Additionally, raising awareness about breast cancer and promoting early detection through regular mammograms and breast self-exams is crucial to improving outcomes and increasing survival rates.

The motivation behind this project is to allow a technique for early and accurate breast cancer detection to prevent unnecessary treatments in cases of false positives and to prevent late treatments due to false negatives.

1.2 Identification of Problem

The goal of this project is to design a deep learning pipeline that can learn how to detect cases of breast cancer in mammograms by detecting the presence of tumour, whether they are benign (non-cancerous) or malignant (cancerous), and their location in the mammogram. The deep learning algorithm will learn the underlying patterns of a large dataset of mammograms to carry out the aforementioned tasks. The motivation behind this project is to allow a technique for early and accurate breast cancer detection to prevent unnecessary treatments in cases of false positives and to prevent late treatments due to false negatives. Ultimately, the target of this project is to combine it with deep learning algorithms developed across other projects.

Fig 1.2: **Breast Cancer**

1.3 Identification of Tasks

To start the project, we have gathered user requirement for this project and prepare the scope and objective. The results from this phase are scope and limitations, objectives, cost, and benefits, feature of the proposed system and user interface design.

This entire project is carried by a group of five individuals. All the phases of this project from proposal till deployment are contributed by all the team members. These phases are as follows:

1. Documentation part
2. Collecting dataset
3. Cleaning data
4. Exploratory data analysis
5. Statistical inference and modelling by applying practical ML
6. Developing and Testing of Data Product
7. Model deployment on website

SPECIAL TECHNICAL TERMS USED IN THIS PROJECT ARE:

1. **Breast Detection**: By using a deep learning algorithm I am going to detect some feature of Breast data and train a ML-model according to that the data that will predict whether a given data of breast is malignant or benign.

2. **Model Deployment:** After training the model I will be going to deploy that model on web, I might change that in future depends on situation.

```
In [20]: x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=5)
         x_train
```

Out[20]:

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst radius	worst texture	worst perimeter	worst area	wo smoothne
306	13.200	15.82	84.07	537.3	0.08511	0.05251	0.001461	0.003261	0.1632	0.05894	...	14.41	20.45	92.00	636.9	0.112
410	11.360	17.57	72.49	399.8	0.08858	0.05313	0.027830	0.021000	0.1601	0.05913	...	13.05	36.32	85.07	521.3	0.145
197	18.080	21.84	117.40	1024.0	0.07371	0.08642	0.110300	0.057780	0.1770	0.05340	...	19.76	24.70	129.10	1228.0	0.088
376	10.570	20.22	70.15	338.3	0.09073	0.16600	0.228000	0.059410	0.2188	0.08450	...	10.85	22.82	76.51	351.9	0.114
244	19.400	23.50	129.10	1155.0	0.10270	0.15580	0.204900	0.088860	0.1978	0.06000	...	21.65	30.53	144.90	1417.0	0.146

Fig 1.3: **Test/Train Dataset**

1.4 Timeline

	WEEK1	WEEK2	WEEK3	WEEK4	WEEK5	WEEK6	WEEK7	WEEK8	WEEK9	WEEK10
PROJECT										
ABSTRACT										
INTRODUCTION										
LITERARTURE REVIEW										
METHEDOLOGY										
MODULE										
PROBLEM DEFINATION										
OBJECTIVE										
BIBIOGRAPHY										

Gantt Chart

1.5 Organization of the Report

The whole report will be classified in five chapters accordingly. These chapters will dealt with every phase of the project "Breast-Cancer-Detection" model. Also the chapters will provide ample knowledge of the objectives that will further divided into the primary and secondary objectives, where primary responds to the early milestones and secondary objectives will accessories them.

- The First chapter will let us know about the initial introduction part of our project where we will be dealing with the stats, documentation, need to resolve the problem, description of problem, identification of problem, benefits, tasks allotted, build and test of solutions.

- The Second chapter will consist writing a complete and extensive literature review about when was the particular problem identified, documentary and the proofs related to it. Also brief of existing model and what are the differences we are going to initiate according to our research. The summary part of the report with project at hand and what problem dealt and how. The review state of deep learning methods applied to the detection of breast cancer in mammograms, which will be used to guide the research towards promising areas and govern the choice of techniques to implement and explore.

- The Third chapter will deal with design and flow of process during the execution and namely features image accretion, image pre-processing, etc. The analysis of features and the selection of design.

- The Fourth chapter will be of the topic results and validation where we will come to know about the implementation of solution using the machine learning and testing part will be observable with the help of diagrams.

- The Fifth chapter will make conclusion of the whole report and also will let know how the particular model/ project is going to benefit the public and also the students for more research in the particular field/topic.

Chapter-2

LITERATURE REVIEW

2.1 Timeline of the reported problem

Breast cancer has been recognized as a disease for thousands of years. The earliest known description of breast cancer dates back to ancient Egyptian times, around 1600 BCE, where it was described in the Edwin Smith Papyrus.

Breast cancer has been recognized and described by many physicians and scientists over the centuries, including the Greek physician Galen, who lived in the 2nd century CE, and the Persian physician Rhazes, who lived in the 10th century CE.

In more recent times, breast cancer was first identified as a distinct disease in the 19th century. In 1857, the Scottish surgeon William Halsted described the radical mastectomy, a surgical procedure that involved removing the entire breast, underlying chest muscles, and lymph nodes.

Today, breast cancer is one of the most common types of cancer worldwide, and early detection and treatment are crucial for improving survival rates. Screening programs and advanced medical technologies have significantly improved the diagnosis and treatment of breast cancer in recent decades.

2.2 Existing Solutions

Over the centuries, various treatments have been proposed for breast cancer, some of which are still in use today.

- **Surgery:** One of the earliest proposed treatments for breast cancer was surgery. In ancient times, breast tumors were often treated with surgical procedures, such as mastectomy (removal of the breast).

- **Radiation therapy:** Radiation therapy was first used to treat breast cancer in the early 20th century. It involves using high-energy radiation to kill cancer cells and shrink tumors.

- **Chemotherapy:** Chemotherapy was first used to treat breast cancer in the 1940s. It involves using drugs to kill cancer cells throughout the body. Today, chemotherapy is often used in combination with other treatments, such as surgery and radiation therapy.
- **Hormone therapy:** Hormone therapy was first used to treat breast cancer in the 1960s. It involves using drugs that block the effects of estrogen and progesterone, which can fuel the growth of some types of breast cancer.

2.3 Literature Review Summary

Year and Citation	Article/Author	Tools/Software	Technique	Source
2015	Breast cancer diagnosis using Genetically Optimized Neural Network model	Machine learning, Deep Learning, Java, Pentium IV computer of 3.4 GHz with 2 GB of RAM	Genetically Optimized Neural Network (GONN) algorithm, for solving classification problems	sciencedirect.com/ science/article/ abs/ pii/ S0957417415000883

2021	Artificial Intelligence (AI) in Breast Imaging: A Scientometric Umbrella Review	CAD, appraisal tools, Measurement Tool to Assess Systematic Reviews, VOS Viewer	Pattern recognition algorithms and deep learning models	mdpi.com/2075-4418/12/12/3111
2011	Analysis of feature selection with classification: breast cancer datasets	Neural Networks, Machine Learning	Decision tree algorithms, Bayesian algorithms, Rule based algorithms	ijcse.com/docs/INDJCSE11-02-05-167.pdf
2020	A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications/ Muhammet Ak	Numpy, Pandas, Matplotlib, Scikit-learn Logistic Regression	K-Nearest Neighbour (KNN), Decision Tree	mdpi.com/2227-9032/8/2/111
2021	An optimized K-Nearest Neighbour based breast cancer detection/ Tsehay Admassu Assegie	Wisconsin breast cancer dataset collected from kaggle data repositor	K-Nearest Neighbour	journal.umy.ac.id/index.php/jrc/article/view/8593
2018	A support vector machine-based ensemble algorithm for breast cancer diagnosis	Weighted Area Under the Receiver Operating Characteristic Curve Ensemble (WAUCE)	support vector machine (SVM)	sciencedirect.com/science/article/abs/pii/S0377221717310810
2014	Performance analysis of support vector machines classifiers in breast cancer mammography recognition	mammography	proximal support vector machine (PSVM), Lagrangian support vector machines (LSVM), Support vector machine (SVM)	link.springer.com/article/10.1007/s00521-012-1324-4

2019	Comparison of the performance of machine learning algorithms in breast cancer screening and detection/Zakia Salod, Yashik Singh	mammography	SVM, BCCD, Adaptive Boosting	journals.sagepub.com/ doi/pdf/10.4081 /jphr.2019.1677
2021	Breast cancer detection based on thermographic images using machine learning and deep learning algorithms/Viswanatha Reddy Allugunti	segmented thermographic images	convolutional neural network (CNN).	Computerscience journals.com /ijecs
2020	Automated Breast Cancer Detection in Digital Mammograms of Various Densities via Deep Learning	mammographic images	Deep learning	mdpi.com/2075-4426/10/4/211

Table 1

2.4 Problem Definition

- Breast cancer can have devastating consequences if not detected and treated early. However, current methods of diagnosis, such as mammography and biopsy, are invasive, costly, and sometimes inaccurate.

- Deep learning techniques could in theory, highly increase the accuracy of mammogram screenings for detecting early signs of breast cancers. However, these techniques require large amounts of data to learn cancer's underlying patterns and adapt to new cases, and require powerful computing resources to accelerate the process of learning the data, making them very hard to optimize.

Chapter-3

DESIGN FLOW/PROCESS

3.1 Evaluation & Selection of Specifications/Features

Based on the machine learning and deep learning applications for the task of breast cancer detection, and the datasets available for this project, design decisions specific to the deep learning implementation will be covered, along with the reasoning behind the choice of datasets to use and general considerations.

- **Dataset Information:** WBDC dataset includes 569 instances with class distribution of 357 benign and 212 malignant. Each sample consists of ID number, diagnosis (B = benign, M = malignant), and 30 features. Features have been computed from a digitized image of a fine needle aspirate of a breast mass Ten real-valued features calculated foreach cell nucleus, and the mean, standard error, and "worst" or largest (mean of the three largest values) of these features were calculated for each image, resulting in 30 features.

- **Independent Component Analysis:** The basic model of ICA is as follows. Suppose that the observed signal is the linear combination of two independently distributed sources also known as sphering data, is the next step. Data which have been whitened are uncorrelated (as PCA). On the other hand, all variables have variances of one. PCA can be used for both these computations because it decorrelates the data and gives information on the variance of the decorrelated data in the form of the eigenvectors.

- **Artificial Neural Networks:** Feedforward neural network (FFNN) is most popular ANN structure due to its simplicity in mathematical analysis and good representational capabilities. FFNN has been used successfully to various applications such as control, signal processing, and pattern classification. FFNN architecture.

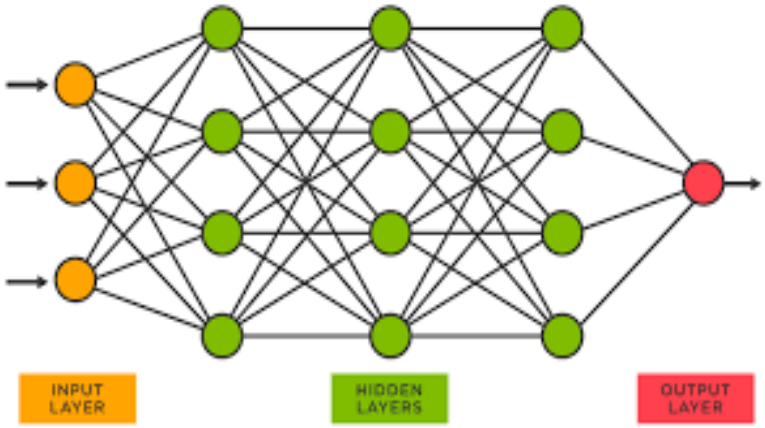


Fig3.1: **Architecture of feedforward neural network**

3.2 Design Constraints

A design constraint is a limitation or requirement that must be taken into consideration during the design process. It can be a physical, functional, or technical limitation that the design must adhere to.

- **Data availability**: The quality and quantity of data available for training the model can significantly affect its performance. It is essential to have a large and diverse dataset that accurately represents the population being studied.

```
{'data': array([[1.799e+01, 1.038e+01, 1.228e+02, ..., 2.654e-01, 4.601e-01,
         1.189e-01],
        [2.057e+01, 1.777e+01, 1.329e+02, ..., 1.860e-01, 2.750e-01,
         8.902e-02],
        [1.969e+01, 2.125e+01, 1.300e+02, ..., 2.430e-01, 3.613e-01,
         8.758e-02],
        ...,
        [1.660e+01, 2.808e+01, 1.083e+02, ..., 1.418e-01, 2.218e-01,
         7.820e-02],
        [2.060e+01, 2.933e+01, 1.401e+02, ..., 2.650e-01, 4.087e-01,
         1.240e-01],
        [7.760e+00, 2.454e+01, 4.792e+01, ..., 0.000e+00, 2.871e-01,
         7.039e-02]]),
 'target': array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
        0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0,
        1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0,
        1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1,
        1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0,
        0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1,
        1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1,
        1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0,
        0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0,
        1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1,
        1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1,
```

Fig3.2: **Scikit Learn**

- **Feature selection:** Identifying the most relevant features that distinguish between malignant and benign tumors is crucial. Too many features can lead to overfit, and few can result in underfit. Therefore, selecting the optimal set of features is a critical design constraint.

```
In [9]: print(cancer['feature_names'])      #features using for classifying the dataset
['mean radius' 'mean texture' 'mean perimeter' 'mean area'
 'mean smoothness' 'mean compactness' 'mean concavity'
 'mean concave points' 'mean symmetry' 'mean fractal dimension'
 'radius error' 'texture error' 'perimeter error' 'area error'
 'smoothness error' 'compactness error' 'concavity error'
 'concave points error' 'symmetry error' 'fractal dimension error'
 'worst radius' 'worst texture' 'worst perimeter' 'worst area'
 'worst smoothness' 'worst compactness' 'worst concavity'
 'worst concave points' 'worst symmetry' 'worst fractal dimension']
```

Fig3.3: **Features Select**

- **Class imbalance:** In breast cancer classification, the number of malignant cases is usually much smaller than the number of benign cases, leading to a class imbalance problem. It is necessary to balance the data or adjust the algorithm to handle the class imbalance problem effectively.

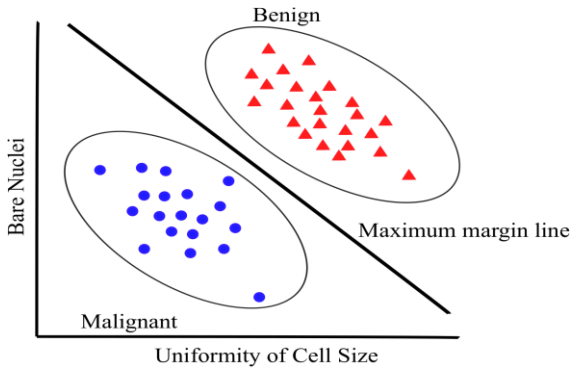


Fig3.4: **Benign and Malignant distinguish**

- **Model interpretability:** SVM is a black-box model, meaning it can be challenging to interpret how the model arrived at its prediction. Interpretable models are often preferred in healthcare settings for transparency and trust. Thus, ensuring model interpretability is another design constraint to consider.

3.3 Analysis of Features and finalization subject to constraints

Our objective is to identify which features are most helpful in predicting malignant or benign cancer and to classify whether the breast cancer is benign or malignant.

Analysis of features and finalization subject to constraints refers to the process of evaluating and selecting features or characteristics of a product, service, or system while taking into consideration any limitations or restrictions that may exist. The process typically involves identifying the features that are desired or necessary for the product, evaluating each feature in terms of its feasibility and value, and then prioritizing and selecting the most important features that can be implemented;

- **Cleaning data:**
  Data quality is one of the most important problems in data management, since dirty data often leads to inaccurate data analytics. Data Analyst spend a large amount of their time cleaning datasets and getting them down to a form with which they can work.
  Check for duplicate data and remove it. Duplicate data can cause the machine learning algorithm to overfit the data and produce unreliable results. They can have a significant impact on the results of the machine learning algorithm, so it is important to remove them.
  There is a dataset named "id" that is of no use for classification, also more than 32+ unnamed features including Nan values and many features which we have no idea about.

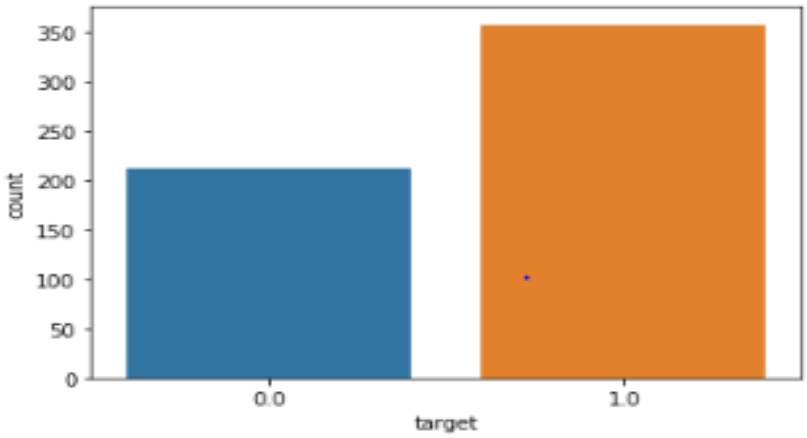


Fig 3.5: **Bar Chart (B and M)**

- **Modifying data:**

  It is the process of figuring out what the data can tell us and we use Exploratory Data Analysis to find patterns, relationships, or anomalies to inform our subsequent analysis. We will look for patterns, differences, and other features that address the questions we are interested in also will try to uncovered the relationships between the different variables.

  For example, in texture, mean feature, median of the Malignant and Benign looks like separated so it can be good for classification.

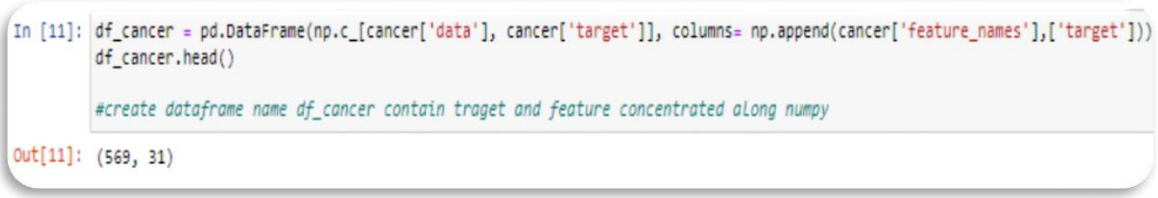


Fig 3.6: **Extract Features**

3.4 Design Flow

The deep learning model aimed to be implemented during this project will require real-life data to learn the underlying patterns in order to be able to detect cases of breast cancer in mammograms and to evaluate its performance. Therefore, the main ethical concern when using sensitive medical data such as mammograms is whether it can be traced back to the original patient. To develop breast cancer prediction models with varying degrees of success there have been two approaches introduced till now; the CNNs and SVMs.

- **CNN (Convolutional Neural Network)**

  Convolutional Neural Networks (CNNs) for Image Analysis, one approach for predicting breast cancer is through the use of medical imaging. A CNN can be trained to analyze mammograms and detect any abnormalities or potential signs of cancer. The CNN can be trained on a dataset of mammograms that have already been diagnosed as cancerous or non-cancerous. Once trained, the model can be used to predict whether new mammograms contain signs of breast cancer. A Convolutional

Neural Network (CNN) is a deep learning algorithm that is primarily used for image classification and object recognition. CNNs are composed of multiple convolutional layers that extract features from an image, followed by pooling layers that down sample the feature maps, and finally, fully connected layers that classify the input image based on the extracted features. CNNs are widely used in applications such as image recognition, computer vision, and natural language processing.

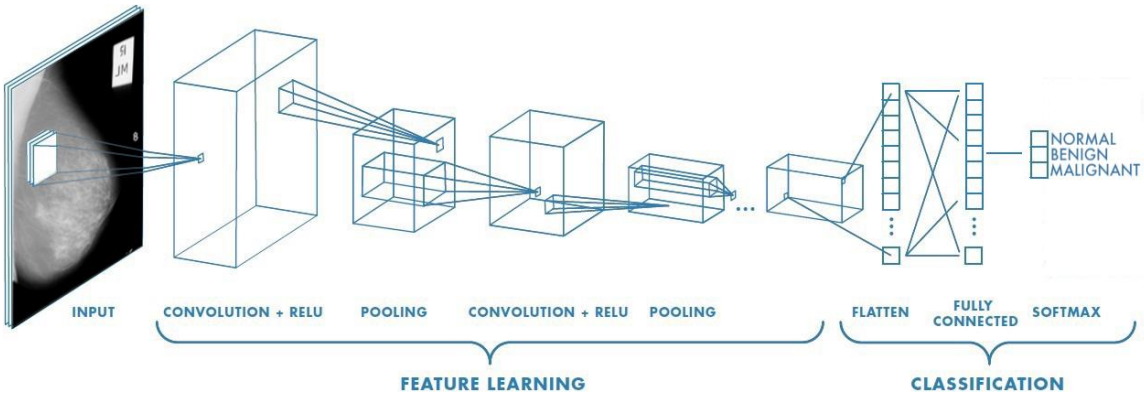


Fig 3.7: **CNN adapted for multi-class breast cancer detection**

- **SVM (Support Vector Machines)**

  Support Vector Machines (SVMs) for Feature Analysis, another approach is to use SVMs for feature analysis. This involves extracting various features from medical images, such as the size and shape of lesions or the texture of surrounding tissue, and then using an SVM to classify the image as either cancerous or non-cancerous. The SVM can be trained on a dataset of images with known diagnoses to learn how to distinguish between the two classes. It is algorithm that can be used for both classification and regression tasks. SVM works by finding the hyperplane that best separates the classes in the feature space. The hyperplane is chosen so that the margin between the hyperplane and the closest points from both classes is maximized. SVMs are effective in high-dimensional spaces and can be used for a variety of applications such as text classification, image classification, and bioinformatics.

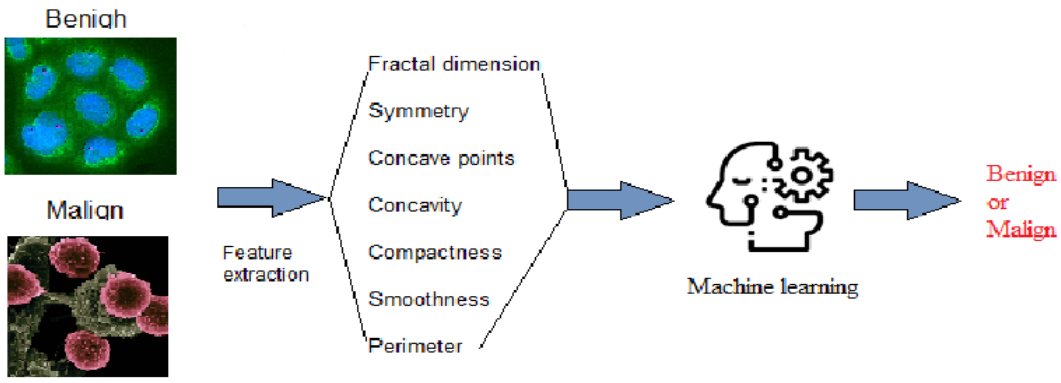


Fig 3.8**: SVM adapted for multi-class breast cancer detection**

3.5 Design Selection

The design selection process for breast cancer detection system should involve careful consideration of several factors, such as data availability, algorithm performance, interpretability, and practicality. Another important consideration in the design selection process is the choice of features to be used in the machine learning model. Feature selection involves identifying the most relevant characteristics of the data that are most predictive of breast cancer. This step can be done manually by experts in the field or using automated feature selection techniques. Through thorough evaluations of these factors, we came to a point of using SVM over CNN.

In this project we are going to use structured data, as numerical data which comes under SVM.

- **Data type:** We are working with structured data, such as categorical data, SVMs here be a better choice. If we were working with unstructured data, such as images or text, CNNs are often a better choice.

- **Data size:** SVMs work well with small to medium-sized datasets. For very large datasets, CNNs may be a better choice because they can handle a larger number of parameters and can learn more complex features.

- **Accuracy:** In general, CNNs are more powerful than SVMs when it comes to accuracy in image recognition and similar tasks. However, SVMs can still be a good choice in some cases where the dataset is small or the features are well-defined.

- **Interpretability:** SVMs are often preferred in cases where the interpretability of the model is important. SVMs produce a hyperplane that separates the data, which can be easily visualized and understood. CNNs, on the other hand, are often considered as "black boxes" because their internal workings are not easily interpretable.

In conclusion, both SVMs and CNNs have their strengths and weaknesses, and the choice of which algorithm to use depends on the specific problem at hand. But for this we are going to use dataset in the form of numerical data which will evaluate whether particular individual have the cancer or not.

SOFTWARE AND HARDWARE REQUIREMENTS:

SOFTWARE REQUIREMENTS:

- WINDOWS 8/9/10
- Python
- Jupyter

HARDWARE REQUIREMENTS:

- RAM:4GB
- Hard disk: 128 GB SSD
- Processor: i5 7th generation
- Display: 1920 X 1080 IP

SOFTWARE:

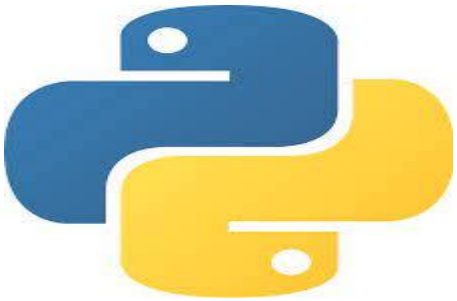


Fig 3.9: **Python**

Fig 3.10**: Jupyter**

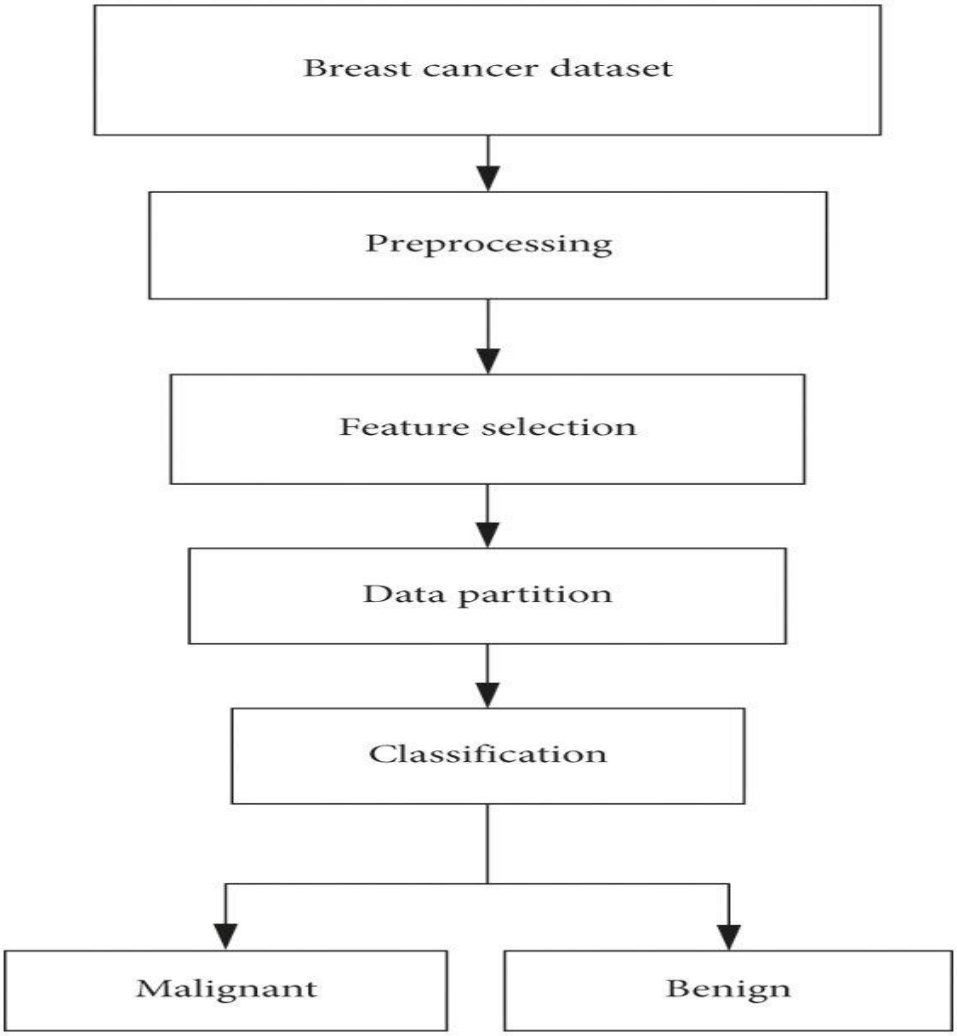


Fig 3.11: **Project work chart**

3.6 Implementation plan/ methodology

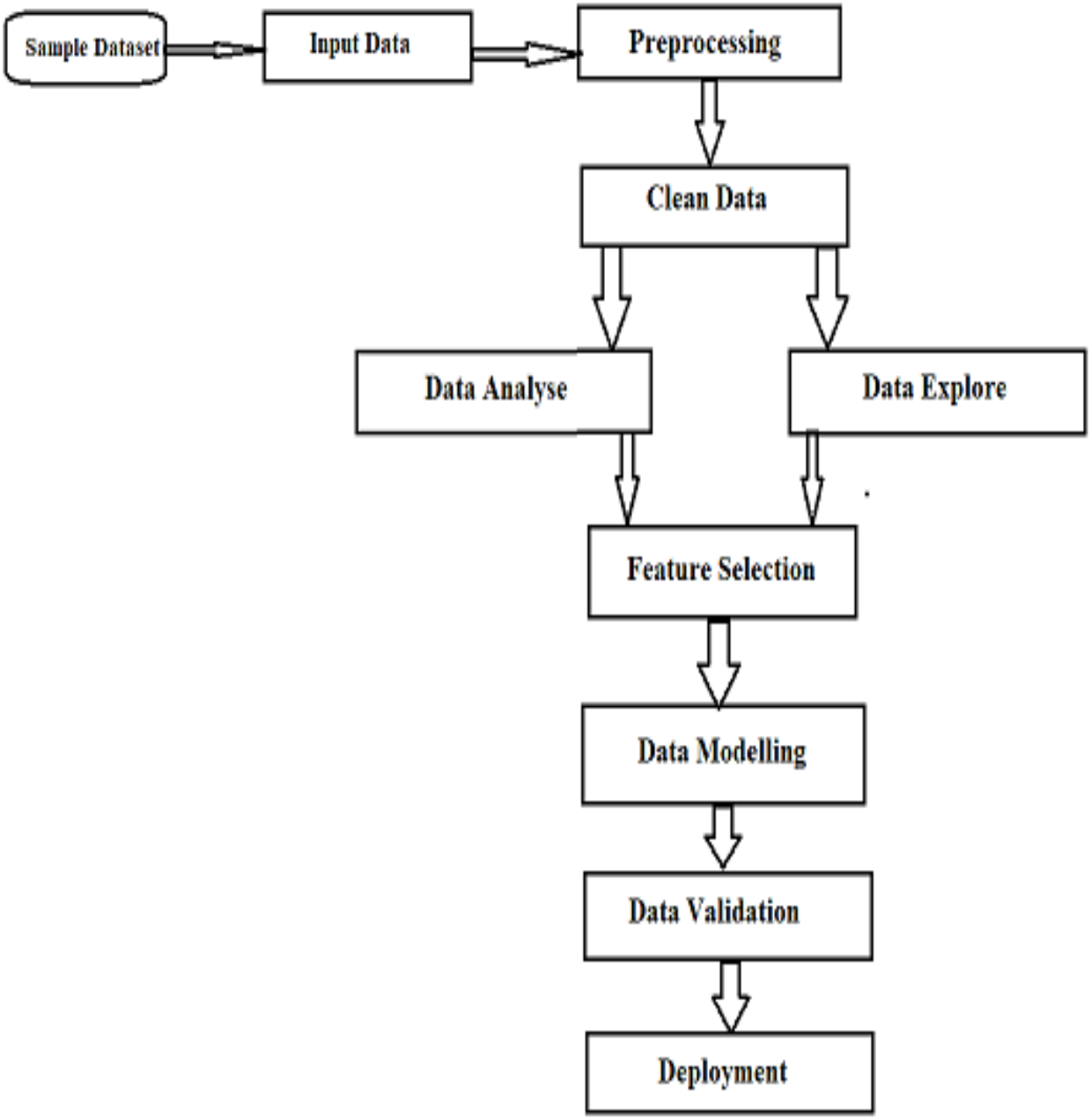


Fig 3.12: **Working model chart**

Chapter-4

RESULTS ANALYSIS AND VALIDATION

4.1 Implementation of Solutions

The Implementation of the project breast cancer detection system was conducted in two parts. The first part corresponds the research-based operation done by group of 3, and the second part to individual extensions and implementations done by group of 2 (software, website).

An overview of both the sections will be observed in this chapter, with work distribution during the implementation of common project.

BREAST CANCER PREDICTION

Using the Breast Cancer Wisconsin (Diagnostic) Database, we can create a classifier that can help diagnose patients and predict the likelihood of a breast cancer.

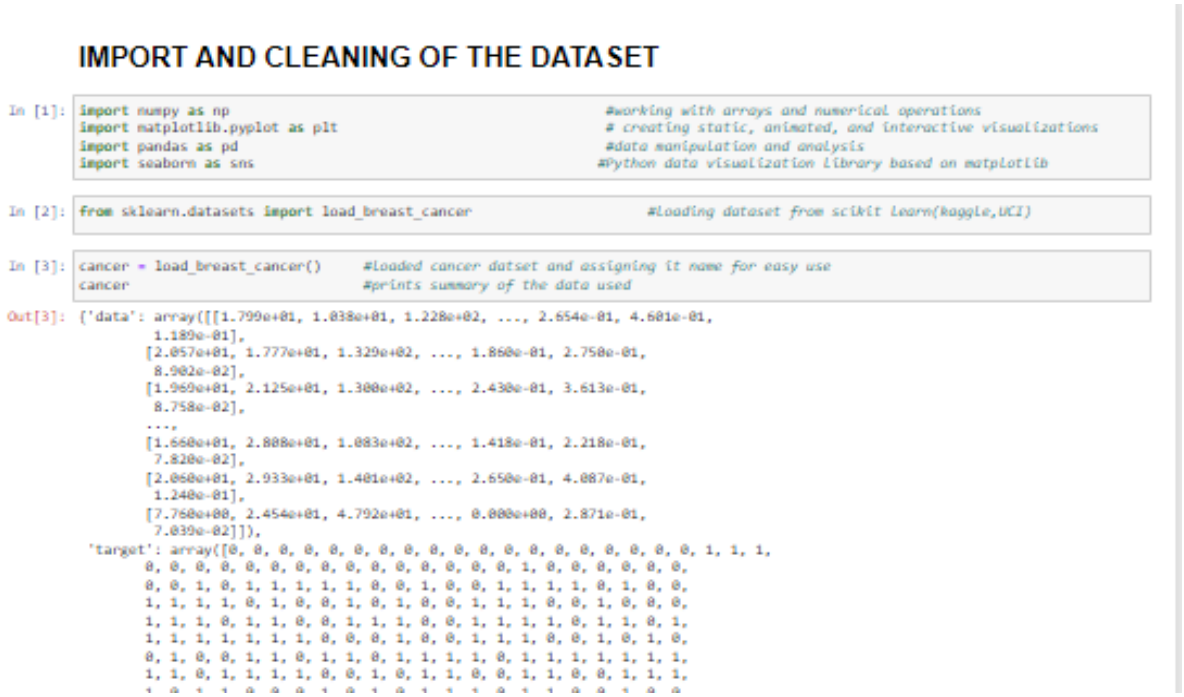
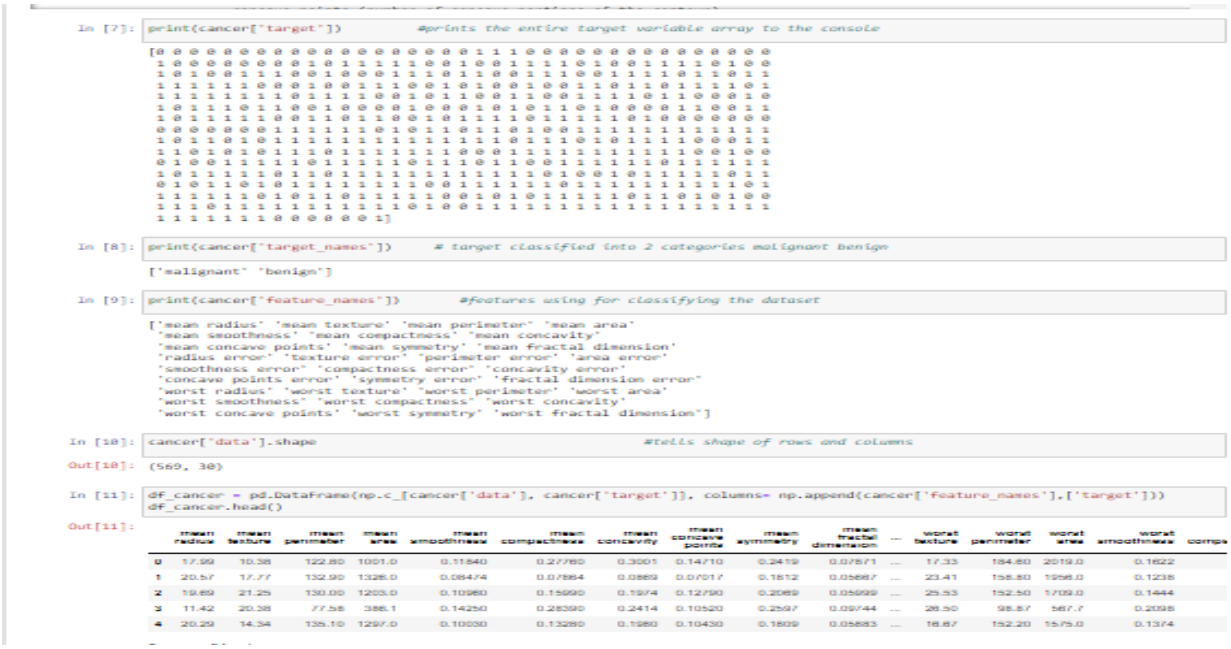


Fig 4.1: **Import of dataset**

Exploratory analysis

Load the dataset and do some quick exploratory analysis. After Inserting the data accessing in order to test images in random from local host.



```
In [7]: print(cancer['target'])          #prints the entire target variable array to the console
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 1 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 0 0 1 0 0 1 1 1 1 0 1 0 0 1 1 1 1 0 1 0 0
 1 0 1 0 0 1 1 1 0 0 1 0 0 0 1 1 1 1 0 1 1 0 0 1 1 1 0 0 1 1 1 1 0 1 1 0 1 1
 1 1 1 1 1 1 1 0 0 0 1 1 1 1 0 1 0 1 0 1 1 0 0 1 0 0 1 1 0 1 1 1 1 1 1 1 0 1
 1 1 1 1 1 1 1 1 1 0 1 1 1 0 0 1 0 1 1 0 0 1 1 0 0 1 1 1 1 0 1 1 0 0 0 1 0
 1 0 1 1 1 0 1 1 0 0 1 0 0 0 0 1 0 0 0 1 0 1 0 1 1 0 1 0 0 0 0 1 1 0 0 1 1
 1 0 1 1 1 1 1 0 0 1 1 0 1 1 0 0 1 0 1 1 1 1 0 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 1 0 1 1 0 1 1 0 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1
 1 0 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 0 1 1 1 1 0 0 0 1 1
 1 1 0 1 0 1 0 1 1 1 0 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0
 0 1 0 0 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 0 1 1 0 0 1 1 1 1 1 1 1 0 1 1 1 1 1
 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 1 1 1 1 0 1 1 0 1 0 1 1
 0 1 0 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1
 1 1 1 1 1 1 0 1 0 1 1 0 1 1 1 1 1 1 1 1 0 0 1 0 1 0 1 1 1 1 1 1 1 0 1 0 1 0
 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 0 0 0 0 0 0 1]

In [8]: print(cancer['target_names'])      # target classified into 2 categories malignant benign
['malignant' 'benign']

In [9]: print(cancer['feature_names'])        #features using for classifying the dataset
['mean radius' 'mean texture' 'mean perimeter' 'mean area'
 'mean smoothness' 'mean compactness' 'mean concavity'
 'mean concave points' 'mean symmetry' 'mean fractal dimension'
 'radius error' 'texture error' 'perimeter error' 'area error'
 'smoothness error' 'compactness error' 'concavity error'
 'concave points error' 'symmetry error' 'fractal dimension error'
 'worst radius' 'worst texture' 'worst perimeter' 'worst area'
 'worst smoothness' 'worst compactness' 'worst concavity'
 'worst concave points' 'worst symmetry' 'worst fractal dimension']

In [10]: cancer['data'].shape                                  #tells shape of rows and columns
Out[10]: (569, 30)

In [11]: df_cancer = pd.DataFrame(np.c_[cancer['data'], cancer['target']], columns= np.append(cancer['feature_names'],['target']))
         df_cancer.head()
Out[11]:
```

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	worst smoothness	comp
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	...	17.33	184.60	2019.0	0.1622	
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	...	23.41	158.80	1956.0	0.1238	
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	...	25.53	152.50	1709.0	0.1444	
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	...	26.50	98.87	567.7	0.2098	
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	...	16.67	152.20	1575.0	0.1374	

Fig4.2: **Cleaning of dataset**

Visualising the dataset

Visualizing a dataset can provide valuable insights and help understand the underlying patterns and relationships within the data. For easy exploration of the dataset.

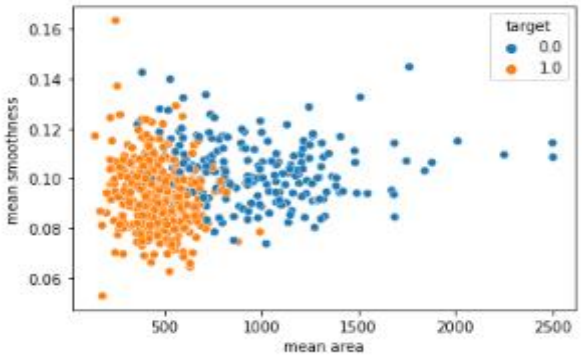


Fig4.3: **Scatterplot**

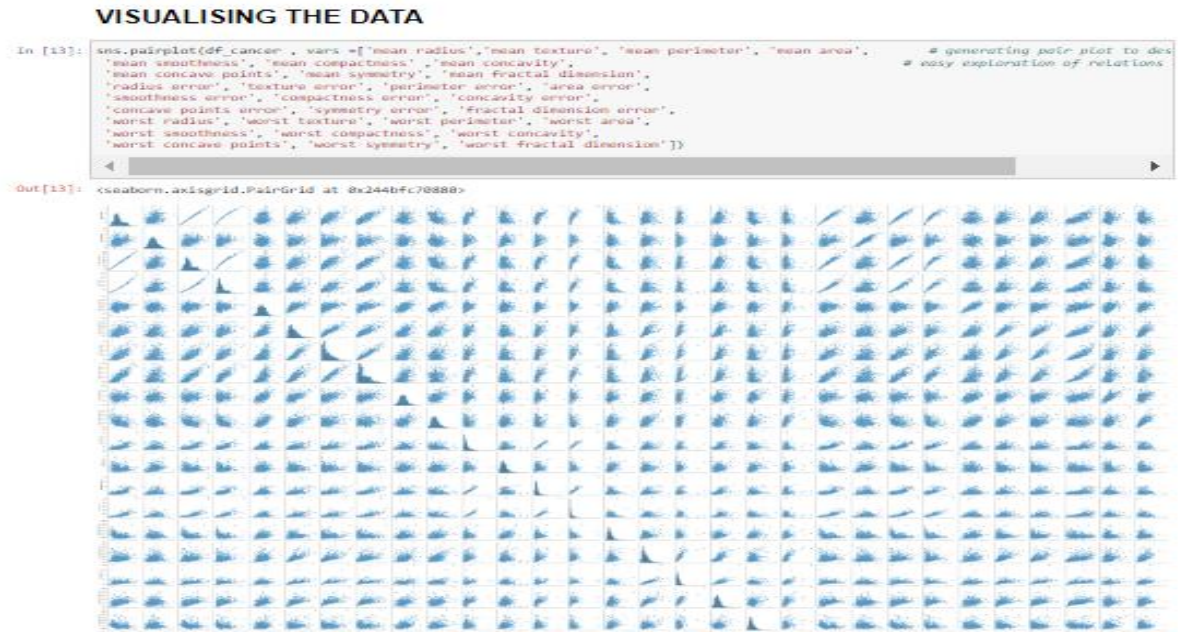


Fig 4.4: **Visualise Dataset (Pair Plot)**

It is good to check the correlations between the attributes. From the output graph below, the red around the diagonal suggests that attributes are correlated with each other. The white and pink patches suggest some moderate correlation and the violet boxes show negative correlations. colour code the cells in the confusion matrix based on the number of observations that fall into each category.

Fig4.5: **Heatmap**

Splitting the dataset

Splitting a dataset involves dividing it into two or more separate subsets for the purpose of training and testing a machine learning model. The most common approach is to split the dataset into training and testing sets, where the training set is used to train the model, and the testing set is used to evaluate its performance.

Using the variables (x_train, x_test, y_train, y_test) the dataset is split into two sets of test and train. Firstly, for any project to be executable we need to ensure from our end i.e. test and will then then train for multi cases.

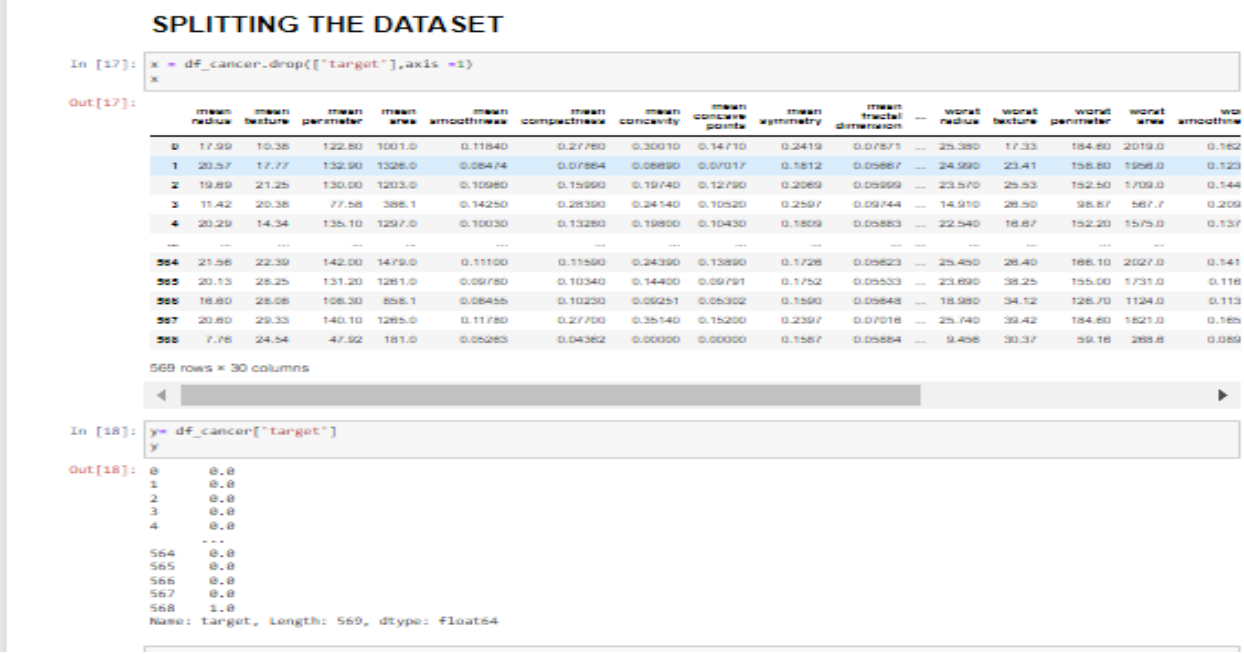


Fig4.6: **Train and Test Dataset**

Training the dataset

Training a machine learning model involves using a dataset to teach the model to make predictions or classify new data based on patterns in the training data.

Here I have trained Support Vector Machine using Scikit-learn library on my dataset. It initializes an SVM classifier using the SVC() class and fits the model to the training data using the fit() method.

TRAINING THE MODEL USING SVM

```
In [30]: from sklearn.svm import SVC

In [31]: from sklearn.metrics import classification_report , confusion_matrix

In [32]: svc_model= SVC()

In [33]: svc_model.fit(x_train,y_train)

Out[33]: SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
           decision_function_shape='ovr', degree=3, gamma='auto', kernel='rbf',
           max_iter=-1, probability=False, random_state=None, shrinking=True,
           tol=0.001, verbose=False)
```

Fig 4.7: **Training Dataset**

Evaluating the dataset

Evaluating a machine learning model is a critical step in the process of developing a successful model. It involves measuring the performance of the model on a dataset that it has not seen before.

After training, the model is used to make predictions on the training set using the predict() method. I used the x_test for the prediction and the predicted table is stored in y_predict. The confusion_matrix() function is used to compute the confusion matrix based on the true labels (y_test) and predicted labels (y_predict).

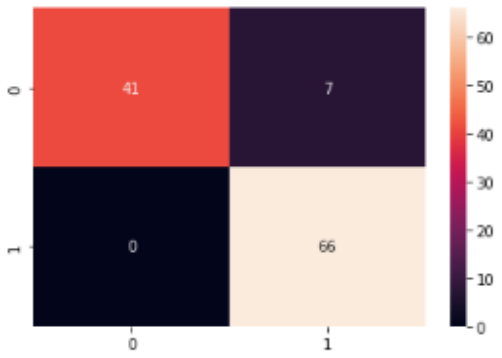


Fig 4.8: **Confusion Matrix**

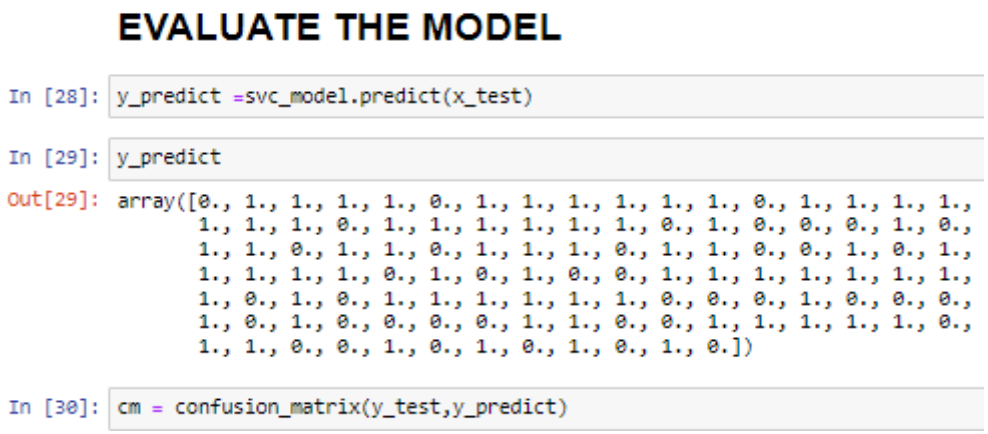


Fig 4.9: **Evaluation of Model**

Accuracy for the dataset

Accuracy is typically used as an evaluation metric for classification tasks. It represents the ratio of correct predictions to the total number of predictions made by a model on a given dataset. However, the accuracy of a dataset alone doesn't make sense without considering the model's performance on that dataset.

The accuracy of the breast cancer model increased by 97% using the modern and new libraries with high features.

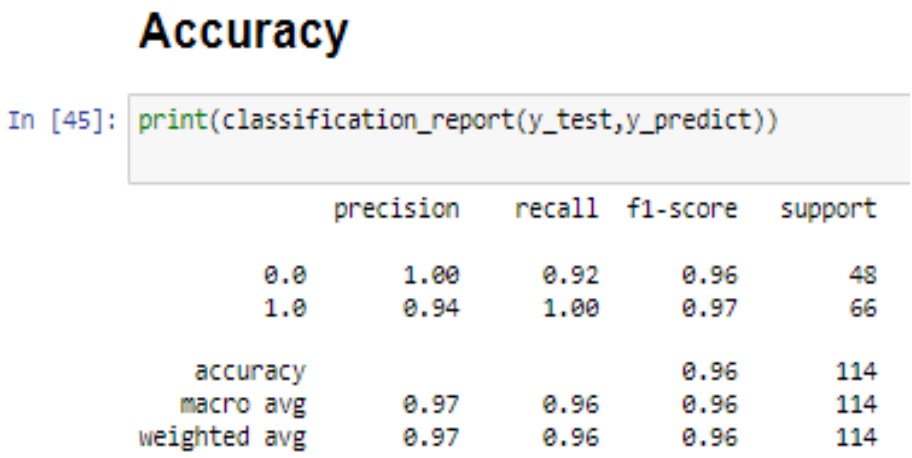


Fig 4.10: **Accuracy**

Chapter-5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

The primary goal of this project was to design and put into action a deep learning pipeline that could identify breast cancer instances in mammograms using a variety of deep learning approaches. A fully-functional pipeline containing data pre-processing processes, a CNN model for learning the data, and prediction visualizations was made after researching a wide range of methodologies utilizing a bag-of-tricks approach.

The concern that predictive analytics may reduce patient care to a set of algorithmically derived probabilities is important and real. Particularly as legislation and governance lags behind technology disruption. However, the benefits are also important and real.

Technology is playing an integral role in the world today and all sectors are benefitting from what it has to offer. The health care sector is no exception. It can benefit significantly from predictive analytics, and it can be argued that this technology is a core aspect of the future of medicine and health care delivery in general.
Diagnosis would be more accurate, as would the treatment that follows. Caregivers would also benefit, given how easy it would be to access useful information and take appropriate steps toward seeing the health of their patients improve.

However, even with these advantages, there are many emerging risks that need to be navigated for all involved parties to benefit from the full potential of predictive analytics. Mostly, this would involve setting clear risk controls to cover bias, address emerging ethical considerations, and ensure clearer documentation for accountability. With policymakers still moving to catch up with the drafting of appropriate legislation, this would require self-regulation from those responsible for writing the algorithms. Existing predictive models and analysis also need to avoid breaking any existing laws such as those around privacy or violating ethical standards.

5.2 Future Work

In the realm of breast cancer detection, the future holds great potential for further advancements in utilizing Support Vector Machines (SVM). SVM has shown promise as a reliable machine learning algorithm for this task. To push the boundaries of breast cancer detection using SVM, future work can focus on several key areas.

The primary area of work that needs improvement is mammogram preprocessing, as this is frequently a region where significant performance gains can be found by using techniques like global contrast normalization (GCN), and local contrast normalization.

Computer vision techniques should be used to eliminate any artifacts, such as tags on the x-rays and black backgrounds, in order to prevent the CNN from picking up useless characteristics.

The fine-tuning of datasets to extract greater performance and prevent over-tatting is another area that can be improved. Images would be smaller with the aforementioned data preprocessing allowing for quicker runtimes, which would allow fine-tuning algorithms like grid search to explore more combinations of configurations in order to uncover better solutions.

This project was a fascinating challenge from my perspective because it had all the traditional difficulties that must be overcome when developing deep learning algorithms, demonstrating unequivocally that developing a high-performance solution is not as simple as it may seem. Having first-hand experience with a family member battling cancer, the chance to apply my expertise to further the field of cancer detection was inspiring.

REFERNECES

- Bhardwaj, A. and Tiwari, A. (2015), `Breast cancer diagnosis using Genetically Optimized Neural Network model', Expert Systems with Applications 42(10), 4611{4620.

- N Houssami, CI Lee, DSM Buist and D Tao,(Sep-2017) propose idea that efforts to improve screening outcomes have mostly focused on large numbers of imaging practices (double instead of single-reading, more frequent screens, or supplemental imaging) that may add substantial resource expenditures and harms associated with population screening.

- F AlFayez, MWA El-Soud and T Gaber - Applied Sciences (Jan-2020) o increase survival rates, it is found that it is very effective to early detect breast cancer. Mammography-based breast cancer screening is the leading technology to achieve this aim. However, it still can’t deal with patients with dense breast nor with tumor size less than 2 mm. Thermography-based breast.

- Zuluaga-Gomez and Z Al Masry (April 2021) A recent study from GLOBOCAN disclosed that during 2018 two million women worldwide had been diagnosed with breast cancer. Currently, mammography, magnetic resonance imaging, ultrasound, and biopsies are the main screening techniques, which require either, expensive devices or personal qualified; but some countries still lack access due to economic, social, or cultural issues.

- D Lavanya and DKU Rani (July-2011) Classification, a data mining task is an effective method to classify the data in the process of Knowledge Data Discovery. A Classification method, Decision tree algorithms are widely used in medical field to classify the medical data for diagnosis. Feature Selection increases the accuracy of the Classifier because it eliminates irrelevant attributes. This paper analyzes the performance of Decision tree classifier-CART with and without feature selection in

terms of accuracy, time to build a model and size of the tree on various Breast Cancer Datasets.

- Li, Y., Wei, Z., Zhao, L., Wang, M., Zhang, Y., & Liu, Z. (2021). A hybrid approach of CNN and decision tree for breast cancer detection using ultrasound images. Computer Methods and Programs in Biomedicine, 207, 106203.

- Ekmekçioğlu, E., Yildirim, O., & Polat, K. (2020). Diagnosis of breast cancer using machine learning algorithms and recursive feature elimination. Journal of Medical Systems, 44(9), 1-8.

- Naseem, I., Akram, T., Qadir, J., & Javed, M. Y. (2019). Breast cancer detection using support vector machine and recursive feature elimination. Journal of X-Ray Science and Technology, 27(5), 755-769.

- Sun, J., Wang, J., Li, Z., Li, J., & Li, Z. (2020). Ensemble of machine learning classifiers for breast cancer diagnosis using mammogram images. IEEE Access, 8, 55635-55646.

- Akram, A., & Hanif, M. (2021). A comparative analysis of machine learning algorithms for breast cancer detection using mammogram images. SN Computer Science, 2(2), 1-15.

- https://github.com/gscdit/Breast-Cancer-Detection

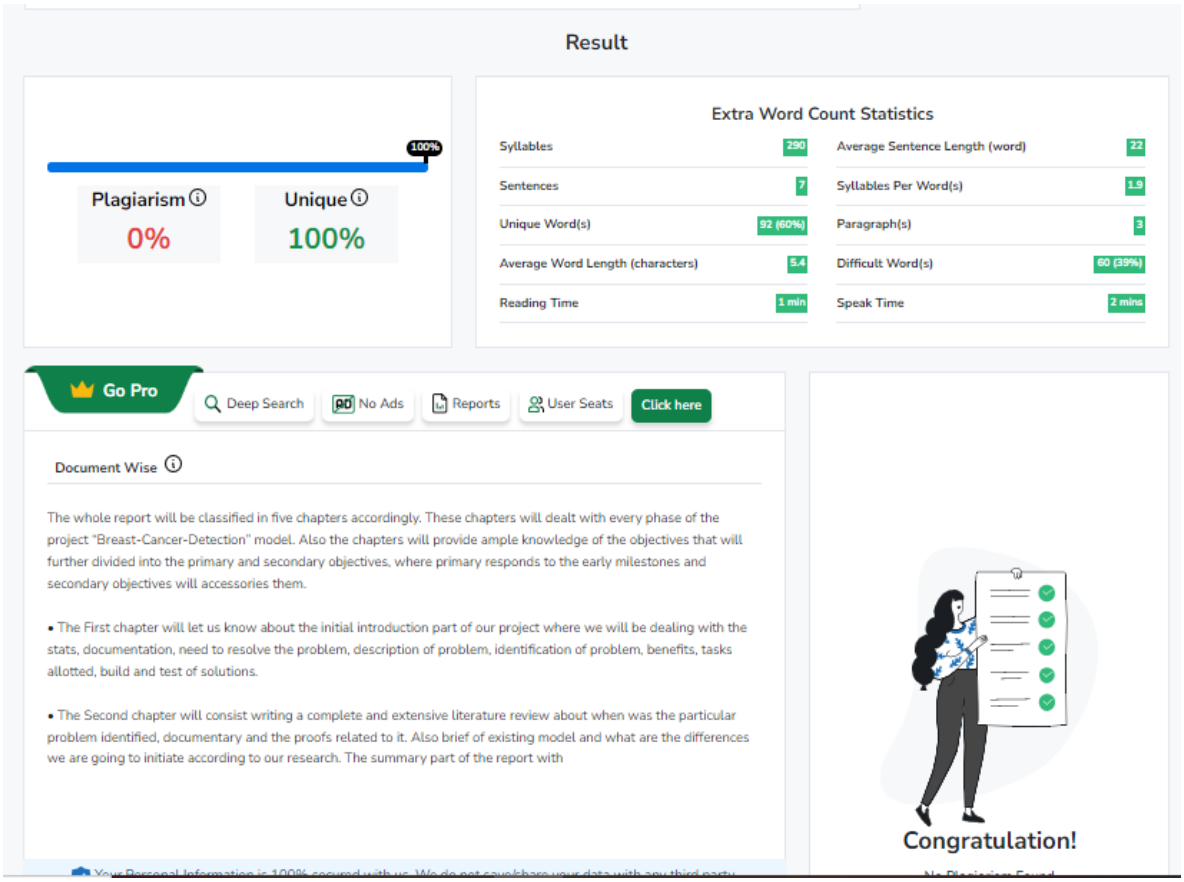
- https://medium.com/analytics-vidhya/building-ml-model-to-predict-whether-the-cancer-is-benign-or-malignant-on-breast-cancer-wisconsin-a09b6c32e7b8

- https://medium.com/@shahid_dhn/building-ml-model-to-predict-whether-the-cancer-is-benign-or-malignant-on-breast-cancer-wisconsin-b8249b55fc62

- https://medium.com/@shahid_dhn/building-ml-model-to-predict-whether-the-cancer-is-benign-or-malignant-on-breast-cancer-wisconsin-d6cf8b47f49a

APPENDIX

Plagiarism Report



Report