LONDON METROPOLITAN UNIVERSITY

islington college
(इस्लिङ्टन कलेज)

CU6051NI - Artificial Intelligence

75% Individual Coursework

Breast Cancer Classification

2022-23

Student Name: Saniya Bist

London Met ID: 22015765

College ID: Np01cp4s220037

Assignment Due Date: Wednesday, January 17, 2024

Assignment Submission Date: Wednesday, January 17, 2024

Word Count: 6500

I confirm that I understand my coursework needs to be submitted online via MySecondTeacher under the relevant module page before the deadline in order for my assignment to be accepted and marked. I am fully aware that late submissions will be treated as non-submission and a marks of zero will be awarded.

Acknowledgements

The satisfaction that accompanies the successful completion of this proposal report would be incomplete without the people who made it possible, without whose constant guidance and encouragement would have made efforts in waste. I consider myself privileged to express gratitude and respect towards all those who guided me through the completion of this report.

I convey thanks to my module leader and tutor, Mr. Rajesh Mahara Sir of module Artificial Intelligence for providing encouragement, guidance, constant support which was a great help to complete this project

Successfully.

Abstract

Machine Learning is the branch of Artificial Intelligence (AI) which focuses on the use of data and algorithms to imitate the human learning process and progressively improve the accuracy. Focusing on the health and applying the algorithms of machine learning over the system that may help patient to get accurate result effectively. This proposal describes on the “Breast Cancer Classification” using the Breast Cancer Wisconsin (Diagnostic) Data set.

For the classification supervised learning model is used with the different algorithms i.e., Support Vector Machine, Decision Trees, Logistic Regression, Naïve Bayes Classifiers. The main objective of the project is to classify the tumors and predict the breast cancer with high accuracy and significant performance. In contrast, its aim is to provide the accurate result and identify its key features analyzing the risk assessment of Breast Cancer Classification

Contents

Table of Figures

1. Introduction

1.1. Explanation of the topic/AI concepts

1.1.1. Artificial Intelligence

Artificial intelligence (AI) is the process of recreating human intelligence in machines that are made to act and think like people. It also describes any machine that displays mental processes like learning and problem-solving. With computers, artificial intelligence seeks to enhance mental processes, perception, and learning. The ideal feature of artificial intelligence is the capacity to reason and make decisions that maximize the likelihood of achieving a given goal.

According to the father of Artificial Intelligence "John McCarthy", it is "*The science and engineering of making intelligent machines, especially intelligent computers programs"* (TutorialsPoint, 2024). The intellectual roots of AI, and the concepts of intelligent machines may be found in Greek mythology. From Turing test to Siri, Alexa to self-driving vehicles AI is progressing rapidly. Though AI is frequently portrayed in science fiction as robots with human-like traits, AI may refer to anything from autonomous weapons to IBM's Watson to Google's search engine (Baum, 2012).

Figure 1: Artificial Intelligence

Although there have been several cycles of hype surrounding artificial intelligence over the years, even critics appear to agree that OpenAI's ChatGPT represents a significant turning point. The first-time generative AI was this significant, advances in computer vision led to breakthroughs; this time, natural language processing is leading the way. Not only can generative models learn the grammar of natural pictures, software code, chemicals, and a wide range of other data kinds, but they can also learn the grammar of language.

Figure 2:- Use of Artificial Intelligence

Our understanding of the potential uses for this technology is still developing, and its applications are expanding daily. However, as the excitement surrounding AI's application in business grows, ethical discussions become increasingly crucial (IBM, 2023).

1.1.2. Machine Learning

Machine learning (ML) is a subset of artificial intelligence (AI). Its main concept is that computer systems may learn on their own using data gathered from previous tasks and experiences. That implies you won't have to pre-program an AI gadget each time you need it to complete a task. Machine Learning is employed in a variety of ways, from automating mundane work to providing insightful insights, and companies across the board are attempting to take advantage. It's possible that you're already using one. For example, a smart home assistant like Google Home or a wearable fitness tracker like Fitbit.

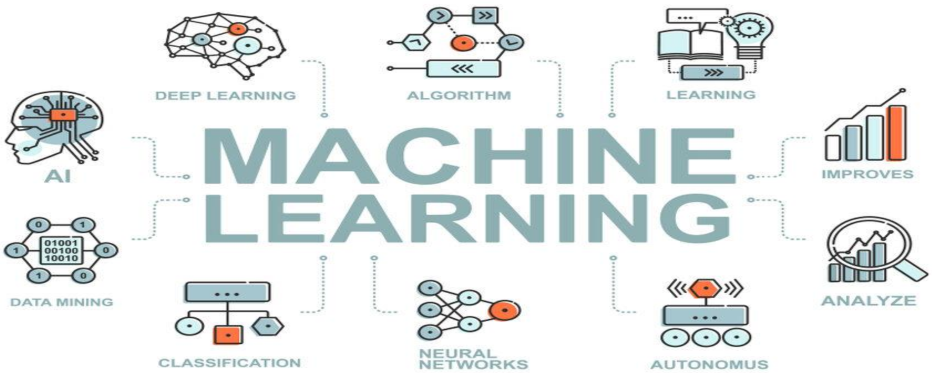

Figure 3: Machine Learning

Deep learning, machine learning, and other subfields are all part of the vast and rapidly expanding subject of artificial intelligence. In a word, machine learning is the idea that computers can use algorithms to learn how to be more creative and predictive so that they can think more like humans (Saha, 2023). Machine learning incorporates several learning techniques, including:

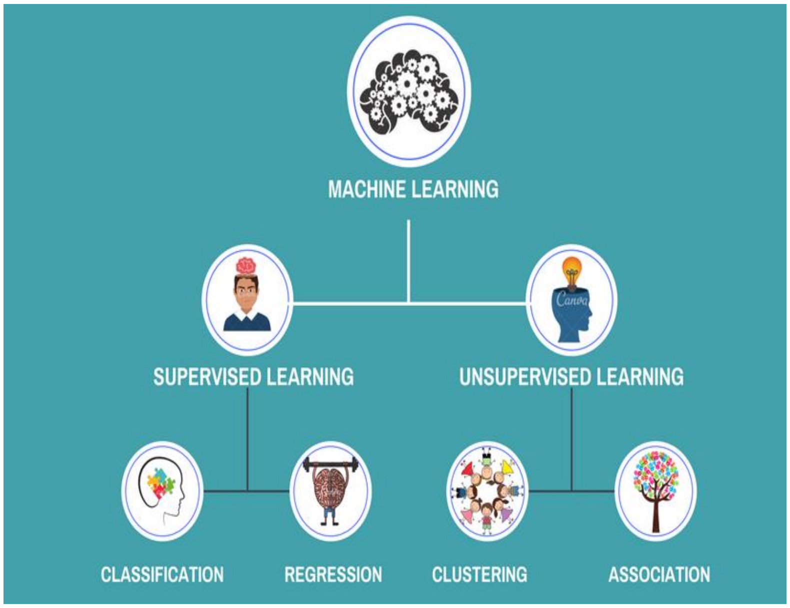

Figure 4: Machine learning types

- Supervised Learning:

  In supervised learning, data with tags is put into our machines, which are then programmed to learn. Our computer is being trained in this process by being given access to a vast quantity of data and taught how to interpret it (Saha, 2023). The types of supervised learning are:

  - Classification
  - Regression

- Unsupervised Learning

  Unsupervised learning is a machine learning approach where users may learn without having supervision over the model. Rather, it let the model do its own thing and find previously missed patterns and information. It mostly addresses the unlabelled data. The types of clustering are:

- Clustering
- Association
- Dimensionality Reduction

- Reinforcement Learning

  A type of machine learning known as reinforcement learning (RL) enables an artificial intelligence (AI)-driven system—also called an agent—to learn by making mistakes and using feedback from its actions. The sole similarity between supervised learning and reinforcement learning (RL) in terms of learning approaches is that both depend on mapping between input and output (Brooks, 2023).

1.2. Explanation/introduction of the chosen problem domain/topic

AI and ML are transforming predictive modelling and diagnostics in the healthcare industry. As one of the main causes of cancer-related fatalities worldwide, lung cancer has emerged as the main target of these technologies. Effective therapy and higher patient survival rates depend on early detection and precise diagnosis.

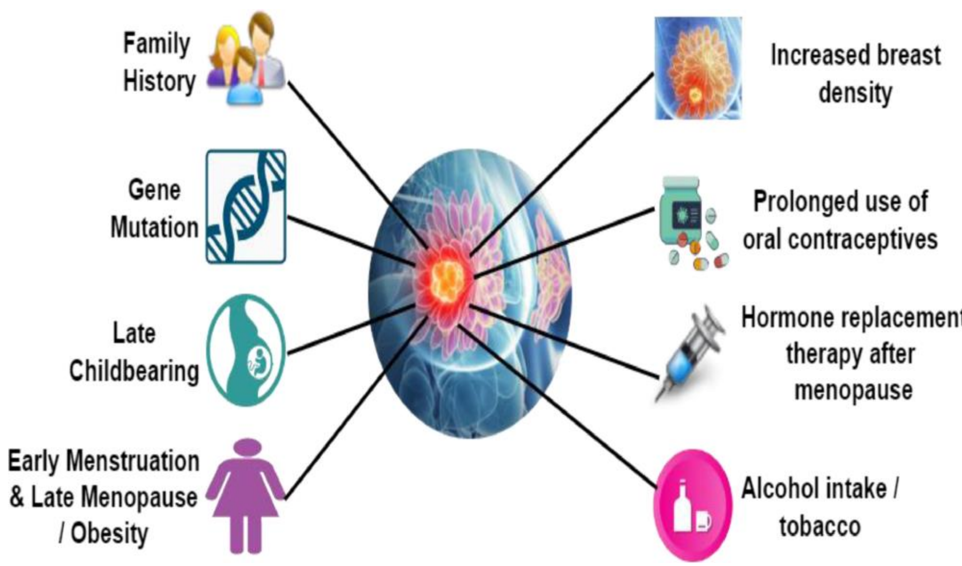

Figure 5: Symptoms of Breast cancer

Breast cancer is one of the main causes of mortality for women worldwide, impacting many of them. It also presents a substantial health burden. Prevention and effective treatment depend on early discovery. Unfortunately, a lot of women are diagnosed with cancer at an advanced stage; breast cancer can originate from fibrous connective tissue, fatty tissues, or breast cells. Different tumours develop in different ways. Some grow quickly, with possibly fatal results. There are two forms of breast cancer: benign and malignant. Risk factors for breast cancer include age and family history. While malignant tumours are more dangerous and can be lethal, benign tumours usually do not represent a threat to life (Godfrey, 2023).

2019 saw the effective deployment of AI by an ICR research team to identify five novel forms of breast cancer. They utilized the trained AI model to identify disease kinds with specific patterns of response to therapy by applying machine learning to molecular data and gene sequences from breast tumours (Godfrey, 2023).

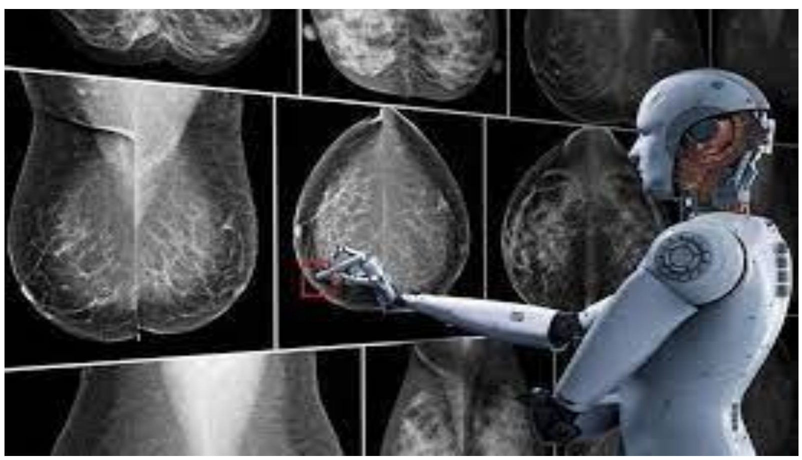

Figure 6: Use of AI for Breast Cancer

In 2020, breast cancer claimed 685,000 lives globally, with nearly half of the cases affecting women without specific risk factors other than gender and age. Approximately 0.5–1% of breast cancer cases occur in men. Various factors, including aging, obesity, alcohol consumption, radiation exposure history, heredity, tobacco use, and postmenopausal hormone therapy, increase the risk of breast cancer. Artificial intelligence (AI) holds immense potential for breast cancer diagnosis and research (IBM, 2023).

The lobules, connective tissue, and ducts are the three primary parts of the breast. Breast cancer typically starts in the lobules or ducts. Breast cancer can be identified by a breast lump or thickening, size, shape, dimpling, redness, pitting, change in nipple appearance, and abnormal nipple discharge. A cancer tumor is a particular kind of cell that develops out of control and infiltrates the tissues around it in the body. Tumors related to breast cancer can be classified as either benign or malignant. A wide range of machine learning (ML), evolutionary machine learning (EML), and deep learning approaches are available for the diagnosis and prediction of breast cancer. Breast cancer diagnosis may be found and predicted using deep learning and machine learning algorithms (Rahul Kumar Yadav, 2023).

2. Background

2.1. Research work done on the chosen problem Domain/ Topic

2.1.1. Breast Cancer Classification

Breast cancer is one of the widely spread disease and can cause death. It has affected around 3,000 pregnant or breast-feeding women. Breast cancer is challenging to detect due to its characteristics as it has rapid growth and later appearance. Nowadays, radiography, X-rays, ultrasounds, are used for diagnosis. The high death rate and widespread burden of breast cancer highlights how important early detection and treatment is (S., 2023).

It is well acknowledged that machine learning and artificial intelligence can solve these problems. One of the main technologies used in medical practice for inspection, diagnosis, and treatment efficiency evaluation is imaging detection, which may indicate differences in tumour size and texture before and after therapy. The huge number of visuals highlights the need for computer-aided identification platforms and techniques, which would free up radiologists' time and allow for a longer reporting period.

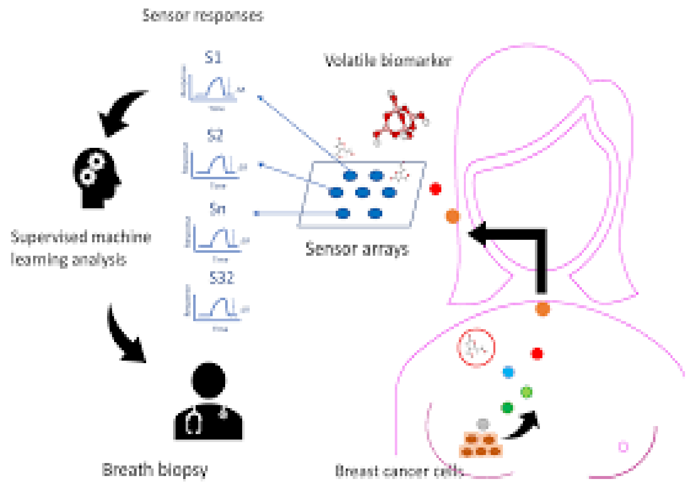

Figure 7: Machine Learning for cancer detection

Developments in the medical field have been facilitated by the incorporation of AI into existing operations. In situations when a physician is unclear or unable of classifying a tumour as benign or malignant, data mining and machine learning techniques may be the crucial part it plays in diagnosis. Early identification of breast cancer by machine learning may result in a longer time to live after receiving a cancer diagnosis. Furthermore, the analysis of patient data and the identification of risk factors made possible by AI and ML have improved breast cancer classification modelling. With the use of these technologies, huge information may be processed, displaying patterns that might not be immediately obvious to human observers and offering information regarding the course and possible recurrence of the disease (Topol, 2019).

2.1.2. Machine Learning to Classify Breast Cancer

The area of artificial intelligence known as "machine learning" deals with computer algorithms that can adapt to new information. It makes use of data-driven technology and computer models to make tasks like categorization, prediction, and detection easier.

In the medical area, machine learning-assisted diagnosis holds the greatest significance. Statistics show that every year in China, about 57 million cases are misdiagnosed to varied degrees, with a misdiagnosis rate as high as 27%. Future advancements in machine learning will lead to ever-increasing accuracy in patient data recognition and imaging diagnosis. This will effectively lower the rate of misdiagnosis and increase diagnosis efficiency, as well as lower the number of medical disputes and offer the public access to high-quality medical care. The burden of treating patients has been decreased because to intelligent medical technologies (Xiao Jia, 2022).

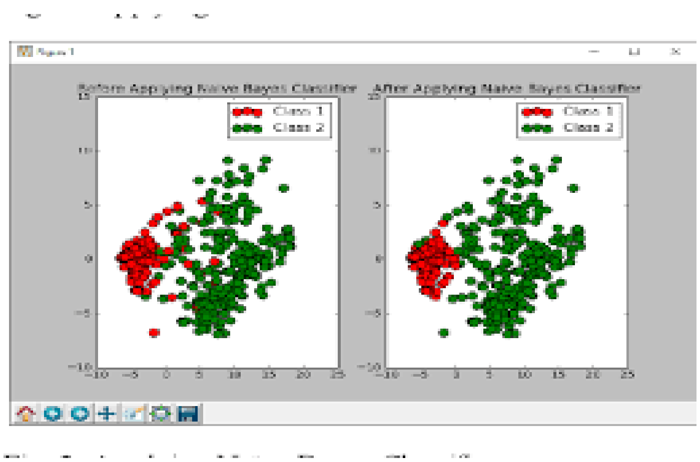

Figure 8: Tumour types

There are several innovative methods to categorize data on breast cancer; some of these include feature selection, while others involve a classification procedure without feature selection. Breast cancer may be identified using Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM) in addition to Naïve Bayes and Weighted Naïve Bayes classifiers for determining the kind of tumours (Khandezamin, 2020).

Triple negative breast cancer and non-triple negative breast cancer patients are classified using a machine learning (ML) technique that uses gene expression data.
When breast cancer is detected, two types of tumours are identified: malignant and benign. The supervised machine learning algorithm can classify it. Furthermore, machine learning models get more precise and efficient when additional information is added to them over time.

In conclusion, machine learning offers a more reliable and approachable technique for early identification and categorization of breast cancer. Using easily accessible data, such medical records, we can identify patterns and produce extremely precise estimations on the presence of malignancies. This might help in the early diagnosis of the condition and enable healthcare providers and individuals to take preventative action.

2.1.3. Advantages of chosen topic/problem domain

In the fields of health care, artificial intelligence and machine learning are often used. In the medical world, it has been highly significant to address diseases early. Artificial intelligence and machine learning have been used in "Breast Cancer Classification" because of its many benefits, which include:

AI and ML has improved accuracy and efficiency, that can learn complex patterns within breast cancer data.

- The machine learning-based technique's major advantage is its capacity to learn on its own from unlabelled raw data. These techniques provide a range of capabilities for breast cancer detection and classification.
- The accuracy of diagnosing breast cancer can now be improved by applying machine learning and image processing techniques for early cancer detection and diagnosis.
- The automation of medical images or patient record analysis with artificial intelligence (AI) reduces the possibility of human mistake, which leads to a more reliable and consistent diagnosis process.
- When it comes to recognizing mammograms that show symptoms of breast cancer in the initial independent read, artificial intelligence (AI) can perform with diagnostic accuracy comparable to that of radiologists.

2.1.4. Disadvantages of Topic/Problem Domain

While classifying and predicting breast cancer using artificial intelligence (AI) and machine learning (ML) has numerous advantages, there are also some serious drawbacks and challenges. They are:

- The technology is not 100% accurate. That may lead to difficulties and serious problems.
- If the models' training data contains biases, the AI system could identify up and enhance such assumptions, producing inaccurate or unfair classification.
- Issues regarding patient privacy, permission, and the proper use of sensitive medical data are among the ethical issues that result from the implementation of AI and ML in healthcare.
- Cyberattacks might target AI and ML systems, and data breaches could expose private medical information.
- The quality, quantity, and accuracy of the training data have a major impact on the way machine learning models perform. The classification of breast cancer may be inaccurate due to incomplete or incorrect information.

2.1.5. Datasets

The dataset for breast cancer classification is made up of a collection of individual health records with various characteristics linked to the risk categorization of breast cancer, including symptoms and signs. The dataset is formatted in a way that facilitates analysis using machine learning algorithms, which search for patterns and connections that may indicate breast cancer. However, several well-known datasets are often utilized for machine learning and data science applications including the categorization of breast cancer. Because of this, one of the most used datasets is the Wisconsin Breast Cancer dataset, also known as the Wisconsin Diagnostic Breast Cancer dataset (Rajamani, 2020).

Attribute Information:

• ID number

• Diagnosis (M = malignant, B = benign)

• Radius (mean of distances from center to points on the perimeter)

• Texture (standard deviation of gray-scale values)

• Perimeter (mean size of the core tumor)

• Area

• Smoothness (local variation in radius lengths)
• Compactness (perimeter^2 / area - 1.0)
• Concavity (severity of concave portions of the contour)
• Concave points (number of concave portions of the contour)
• Symmetry
• Fractal dimension ("coastline approximation" – 1)

2.2. Review and analysis of work in problem domain

2.2.1. Breast Cancer Classification Using Logistic Regression

Authors: Viswanatha V, Ramachandra A.C, Avinash Bhagat and Shashank Shekhar **Journal**: High Technology Letters, ISSN NO: 1006-6748

In this study, the use of deep learning—a branch of machine learning—to develop a breast cancer classification model is addressed. This study used a logistic regression model to classify breast cancer as harmless or cancerous. The dataset utilized is the Wisconsin Diagnostic Breast Cancer dataset, which contains a wide range of factors related to tumours characteristics. The dataset is examined and shown before being converted into training and testing sets. A logistic regression model is trained and assessed using accuracy measures. In the end, the trained model is used to forecast the likelihood that a breast tumour would be malignant. This study highlights how important it is to appropriately categorize breast cancer and demonstrates the value of logistic regression in accomplishing this goal (V, 2023).

2.2.2. Breast Cancer Detection and Classification Empowered with Transfer Learning

Authors: Sahar Arooj, Atta-ur-Rahman, Muhammad Zubair, Kalid Alissa
Journal: Front Public Health

Breast cancer is a specific kind of cancer that starts in the breast and spreads to other parts of the body. This study report suggests an approach for two forms of breast cancer: benign and malignant. Utilizing ultrasonography pictures and Breast cancer histology, identification, and classification are essential steps in computer-aided diagnostic systems. Researchers have demonstrated that it is feasible to automate the fundamental stages of tumour identification and categorization throughout the last few decades. Deep learning (DL), machine learning (ML), and transfer learning (TL) techniques are applied to tackle a lot of medical issues. A number of limited scientific studies on the identification and categorization of cancerous tumours using various models have been released in the past. However, research is challenging when there is no dataset. The proposed method has the goal to support automated breast cancer detection and diagnosis (Arooj, 2022).

2.2.3. Breast Cancer Classification and Prediction using Machine Learning

Authors: Nikita Rane, Jean Sunny, Rucha Kanade, Prof. Sulochana Devi

Journal: International Journal of Engineering Research & Technology (IJERT)

In this paper, the authors present a thorough exam (Rane, 2020)ination of various machine learning algorithms and their application in Breast cancer classification and prediction using Machine Learning. The researchers utilize the different methodologies such as data mining, algorithms, blood analysis, etc. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which was obtained from an MRI image that has been digitally transformed, is used in the present research to compare six machine learning (ML) algorithms: Naive Bayes (NB), Random Forest (RT), Artificial Neural Networks (ANN), Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision Tree (DT). The dataset was divided into two phases: training and testing, to apply the ML algorithms. The model will then categorize the cancer as benign or malignant, and the algorithm with the best results can be utilized as the website's backend. The primary goal of this research is to investigate the performance of classification algorithms to make an early diagnosis of breast cancer (Rane, 2020).

2.2.4. Early Detection of Breast Cancer Using Machine Learning

Authors: Wasi Mohammad Fuad

Journal: BACHELOR OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

The goal of this research is to create a Computer-Aided Diagnosis (CAD) system that can accurately and efficiently identify breast tumours that are malignant. Numerous machine learning methods are used in the study, both with and without Principal Component Analysis (PCA), such as Random Forest, Support Vector Machine, K-Nearest Neighbours, Naïve Bayes, and Logistic Regression. Furthermore, deep learning models are used, including Convolutional and Artificial Neural Networks. Deep learning models often outperforms other classifiers, according to the comparison analysis, with outstanding outcomes in measures like accuracy (98.83%), precision (98.44%), and recall (100%). In the area of predicting breast cancer tumours using digital pictures from fine needle aspirates (FNAs), results generally point to the superiority of deep learning (Faud, 2018).

2.2.5. Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms

Authors: Sharmin Ara, Annesha Das, Ashim Dey

Journal: International Conference on Artificial Intelligence (ICAI)

This research project focuses on breast cancer, which is the most frequent disease in women diagnosed worldwide and has a chance of recovery if caught early. Recognizing the difference between benign and malignant tumours is essential for successful treatment. The research uses machine learning algorithms—Support Vector Machine, Logistic Regression, K-Nearest Neighbours, Decision Tree, Naive Bayes, and Random Forest—to predict breast cancer using the Wisconsin Breast Cancer Dataset, which is obtained from the UCI repository. With an accuracy of 96.5%, performance evaluation shows that Random Forest and Support Vector Machine perform better than other classifiers. The outcomes of this study illustrate to the possible application of these algorithms in the development of an automated diagnostic system for the early diagnosis of breast cancer (Dey, 2021).

2.2.6. Summarized Review and Analysis

In the summarization the collective review and analysis of the research paper, uses variety of predicting, classifying machine learning algorithm for the Breasts Cancer Prediction.

According to the “Breasts Cancer Classification Using Logistic Regression” research paper it mainly focuses on the logistic regression for breast cancer classification and uses Wisconsin Diagnostic Breast Cancer dataset. It emphasizes the importance of accurate and effective classification using logistic regression.

“Breast Cancer Detection and Classification Empowered with Transfer Learning” addresses breasts cancer detection and classification using deep learning and transfer learning and contribute to automated detection and diagnosis. Similarly, “Breast Cancer Classification and Prediction using Machine Learning” examines various machine learning algorithm for breasts cancer classifications. It compares algorithm using Wisconsin Diagnostic Breast Cancer dataset, dividing it into train and test phases, and focuses to implement best performing dataset.

"Early Detection of Breast Cancer Using Machine Learning" aims to create a Computer-Aided Diagnostic system for breast cancer identification. It utilizes different machine learning methods, emphasizing deep learning models that performs accuracy, precision, and recall. Lastly, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms" focuses on distinguishing between benign and malignant tumors using different machine learning algorithms. It uses Support Vector Machine, Logistic Regression, and Wisconsin Diagnostic Breast Cancer dataset.

In overview, these research highlights a diverse range of techniques and strategies for classification of breast cancer from deep learning and algorithms to image processing and machine learning. They highlight the ongoing demand for innovative techniques in early detection and classification for the diagnosis breast cancer.

3. Solution

3.1. Elaboration of used Algorithm (Logistic Regression Model)

There are four algorithms that are mostly used in the Breast Cancer Classification - Support Vector Classifiers (SVC), Logistic Regression, Decision tress, Naïve Bayes Classifiers, each of them has unique features and approaches.

- Support Vector Machine

The support vector machine (SVM) is a machine learning algorithm that tackles classification, regression, and outlier detection by finding the best way to separate data points into different groups. SVMs are used in various fields of application like healthcare, natural language processing, signal processing applications, and speech & image recognition fields.

There are two main categories of support vector machines:

- Simple or linear SVM
- Kernel or non-linear SVM

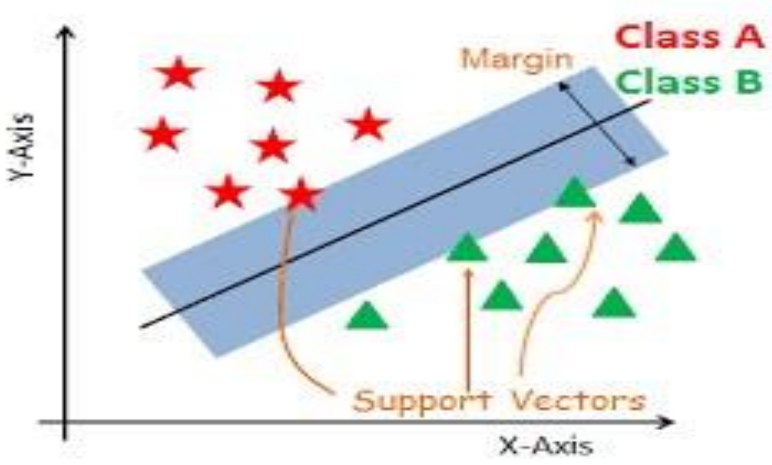

Figure 9: Support Vector Machines

Some of the examples of SMVs are Addressing the geo-sounding problem, Assessing seismic liquefaction potential, Protein remote homology detection, Data classification, Facial detection & expression classification and so on (Kanade, 2022). In two- dimensional space, the equation of the hyperplane is given by:

$$\mathcal{F}(\chi) = w.x + b$$

Equation 1: Support Vector Machines in two-dimensional space

Here,

F(x): decision function that determines the class of the input x.

W: weigh vector, perpendicular to hyperplane

X: input feature vector

B: bias term

- Decision Trees (DTs)

The irregular supervised learning technique for regression and classification is called a decision tree (DT). The objective is to build a model that, by utilizing basic decision rules derived from the data characteristics, predicts the value of a target variable. A component-wise constant approximation can be thought of as a tree (learn, 2023). It can perform multi-class classification on a dataset. According to this method, decisions branch out and lead to consequences, creating a structure or visualization like a tree. In machine learning, decision trees are a method for organizing the algorithm. The cost function will be used to divide the dataset features using a decision tree approach.

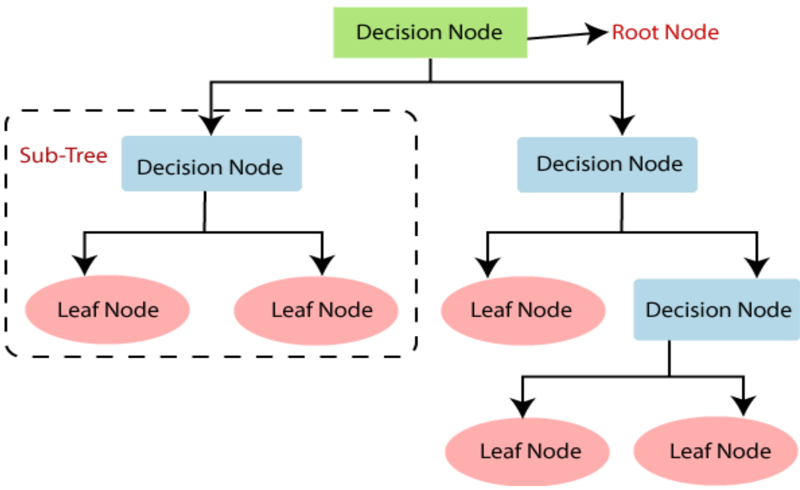

Figure 10: Decision Tree

- Logistic Regression

Logistic regression is the method of modelling the probability of a discrete result given an input variable. The most popular logistic regression models represent a binary result, which is one that can have two possible values, such as true/false, yes/no, and so on. It simulates the probability by allocating values between 0 and 1 using a sigmoid function. This method is beneficial in several industries, including marketing, finance, and healthcare, as it sets a decision limit based on input attributes. Logistic regression is still a fundamental method in machine learning for solving binary classification problems because of its interpretability and simplicity (Thomas W. Edgar, 2017).

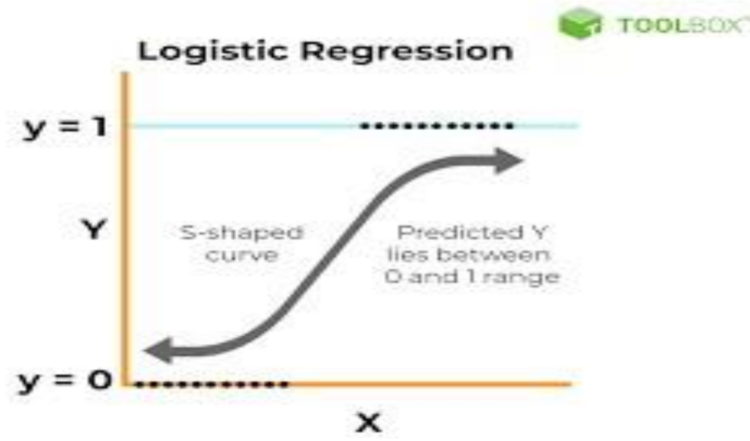

Figure 11: Logistic Regression

The logistic function is defined as:

$$p(Y = 1) = 1\frac{1}{1 + e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n}}$$

Equation 2: Logistic Regression Equation

Here,

P (Y= 1) : probability that the dependent variable Y is 1

e: base of the natural logarithm

B_0: intercept term

B_1, B_2, ... B_n: coefficient associated with independent variables X_1, X_2,….X_n respectively.

• Naïve Bayes

The Naive Bayes classifier is one well-liked supervised machine learning approach for classification tasks, including text categorization. It is a member of the generative learning algorithm family, which implies that it simulates the input distribution for a certain class or category. This method relies on the presumption that the input data and attributes are conditionally independent given the class, which enables the algorithm to provide precise and fast predictions (Ray, 2023). Consider that conditional probabilities, which are expressed by the following formula, indicate the probability of an event given the occurrence of another event.

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)}$$

Equation 3: Naive Bayes equation

Here,

P(A): probability of A occurring.

P(B): probability of B occurring.

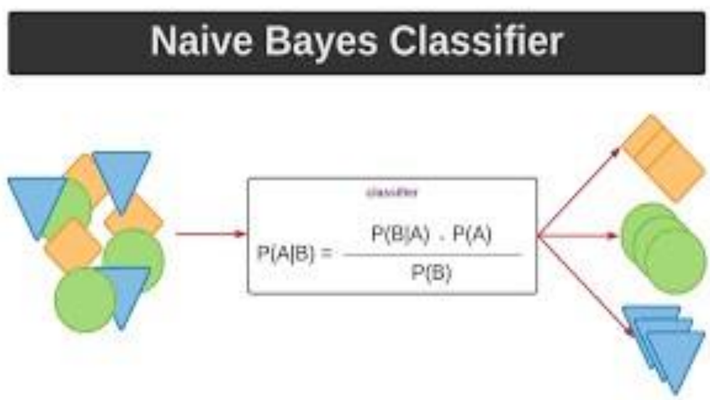

Figure 12: Naive Bayes Classifier

3.2. Implementation methodology of used Algorithm in the system.

The implementation of the methodology of the algorithm used in this system are – Support Vector Classifiers (SVC), Logistic Regression, Decision tress, Naïve Bayes Classifiers, to design Breast Cancer Classification. Preprocessing is used to raw data, which contains characteristics like tumor size, cell shape, and other important characteristics. To maintain consistency and algorithm compatibility, this involves solving missing values, normalizing data, and storing variable types.

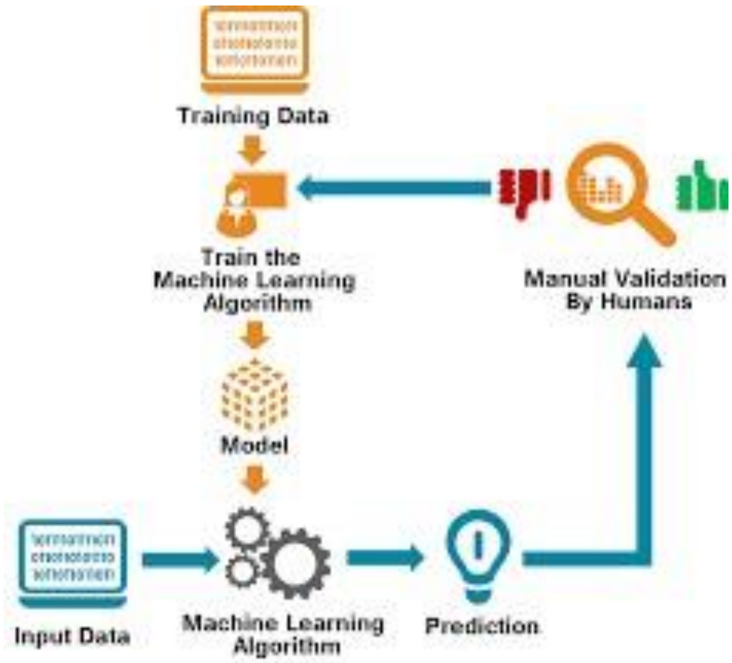

Figure 13: Machine Learning model for classification

Similarly, the dataset is divided into training and testing sets. The chosen algorithm is trained using labeled data. It creates the patterns and the relationship within data. The training set is used to build and train the models, while the testing set is reserved for evaluating their performance. They are also tested on the unseen testing data to evaluate their performance, accuracy, effectiveness, sensitivity in classification of breast cancer. The conclusions that result of these tests provide information about how well each model functions and can direct further optimization and fine-tuning.

Hence, the algorithm which performs most effectively for the specific task of detecting breast cancer is chosen based on its analysis results. Finally, the selected model is integrated into the system. The breast cancer classification system may use the advantages of SVC, Naive Bayes, Logistic Regression, and Decision Tree algorithms to produce accurate and reliable medical diagnosis predictions by using this approach.

3.3. Explanation of the solution/used AI algorithm.

An algorithm is a process or technique in programming that is used to solve an issue. It is based on performing a set of predetermined steps that explain how to accomplish a task; your computer will always carry out these activities properly (Team, 2023). The algorithm for the project is given:

Step 1: START

Step 2: Import necessary libraries,

Step 3: Loading datasets,

Step 4: Data Cleaning,

Step 5: Exploratory Data Analysis,

Step 6: Statistical Analysis,

Step 7: TRAIN and TEST model,

Step 8: Apply Machine learning algorithms,

Step 9: Model Evaluation,

Step 10: Classify Best Models,

Step 11: END

3.4. Pseudocode

Pseudocode is an informal method of describing programming that doesn't depend on underlying technological factors or formal programming language syntax. It helps in the preparation of program rough drafts and/or outlines. A program's flow is summarized using pseudocode, but the underlying details are left out (Coleman, 2024).

The pseudocode for the proposed solution is as follows:

START

IMPORT necessary libraries

LOAD dataset

IMPORT dataset

GET dataset INTO data shape

IDENTIFY and REMOVE duplicate errors

HANDLE missing values

ELSE

IF ANY missing values exist

CALCULATE appropriate statistics(e.g., mean, median) APPLY values TO missing values

ENDIF

PREPROCESS dataset

NORMAIZE or STANDARDIZE numerical variables

SPLIT dataset INTO train, test set

DEFINE models parameters

TRAIN each model on training set

TEST each model on testing set

CHECK performance metrics such as accuracy, precision of each model in training and testing data

CALCULATE value as Benign or Malignant

DISPLAY classification system

GET OUTPUT

SAVE model for future use

END

3.5. Flowchart

A flowchart is a graphic representation of a computer program, system, or process. They are extensively used in many different sectors to plan, analyze, document, and communicate—in simple, understandable diagrams—often complicated processes (LucidChart, 2024).

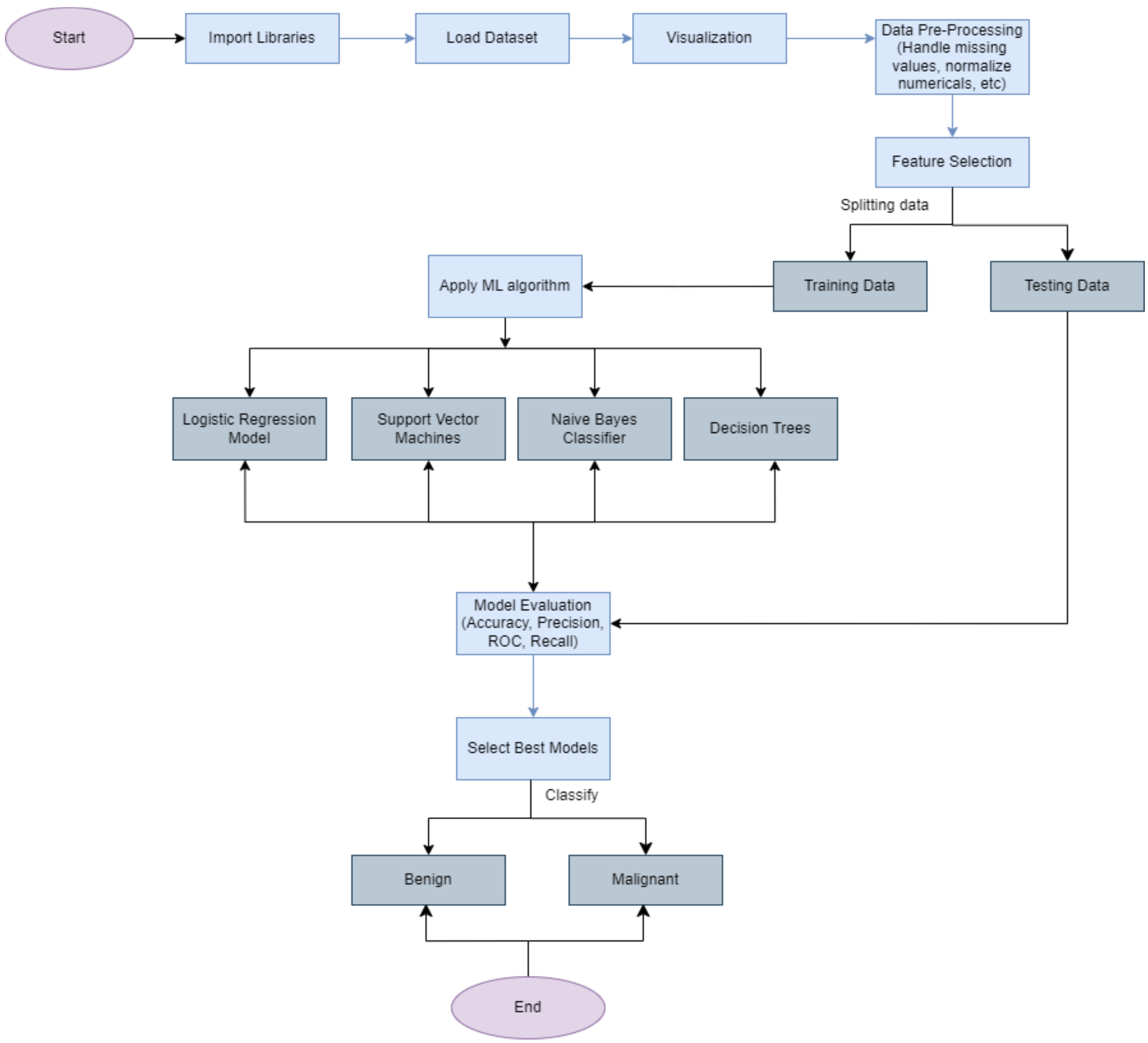

Figure 14: Flowchart

3.6. Development Process

The development process is the implementation of a classification system using datasets and different machine learning algorithms. It begins with importing necessary libraries like pandas, NumPy, seaborn, scikit learn, sklearn etc. Loading the Breast Cancer datasets that contain the information regarding the tumour. Performing the exploration data analysis by understanding the structure, distribution, and characteristics of datasets. Identifying missing values and patterns. Enhanicng datasets by creating new features, label to improve model performance.

After that, train test split process takes place. Where, dataset is named as Breast_Cancer_df is divided into test data set and train data set. Evaluate and select machine learning algorithms suitable for breast cancer classification. Commonly used algorithms like Support Vector Classifiers, Logistic Regression, Naïve Bayes Classifiers, and Decision Trees Classifiers are considered.

Evaluating the model using appropriate metrics such as accuracy, precision, recall, and F1-score. Compare the performance of different models to choose the most effective one for breast cancer classification. Deploy the selected model, integrating it with existing healthcare systems for seamless application.

3.7. Tools and technology

3.7.1. Programming Language

- Python: Python is a high-level, interpreted, object-oriented language with dynamic semantics. The capacity of Python to handle packages and modules makes it easier to create modular applications and reuse code. Since Python is an open-source community language, many self-employed programmers are always creating modules and features for it. Data analysis, scientific computing, artificial intelligence, and backend web development are all greatly enhanced by Python. Python is also used by developers to create desktop programs, games, and productivity solutions (teradata, 2024).

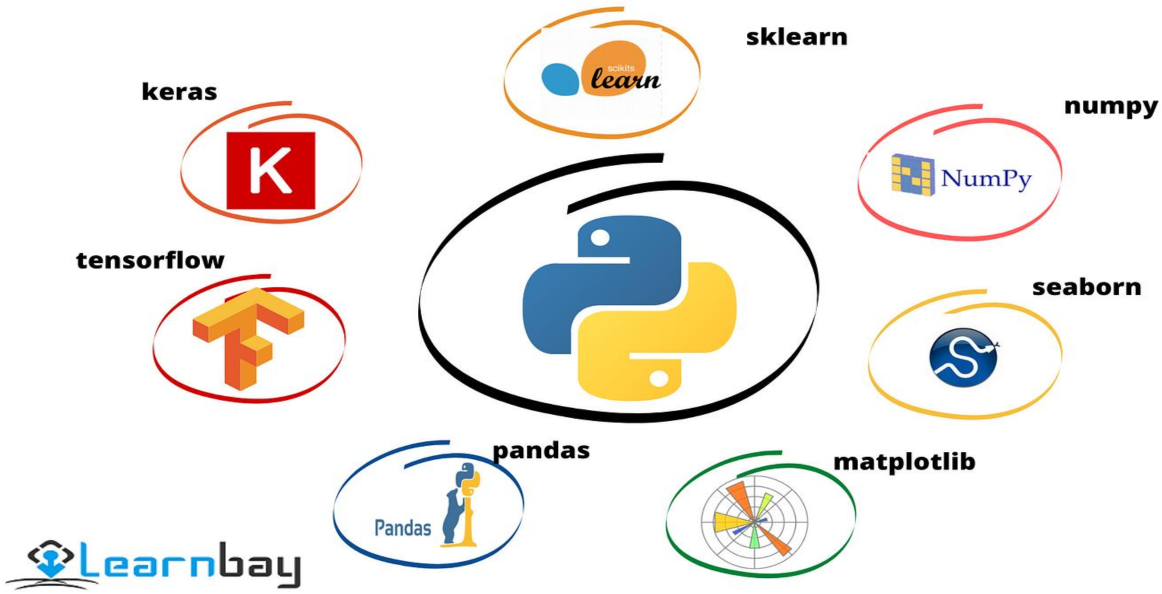

Figure 15: Used tools and technologies

3.7.2. Libraries used

- Pandas: Pandas is a widely used library for working with data it has achieved its objective of becoming the most stable and versatile open-source tool. Pandas focus upon Data Frames. It is easy to modify and work with a panda DataFrame. Pandas provides useful utilities to handle missing values, manipulate data, and conduct operations on rows and columns. As if that weren't enough, several SQL operations like join, merge, filter by, and group by also have panda's equivalents. Pandas is a very popular tool among data scientists, that isn't unusual with all these strong features (Codeacedemy, 2023).

- NumPy: An open-source Python toolkit called NumPy makes it easier to perform effective numerical operations on massive amounts of data. A few of NumPy's methods are utilized with pandas Data Frames. For us, the fact that pandas are built on top of NumPy is its most significant feature. Therefore, NumPy depends on Pandas (Codeacedemy, 2023).

- Matplotlib: Matplotlib is a well-known Python data visualization package. Python visualizations that are static, interactive, and animated are frequently created with it. using only a few lines of code, you can create plots, histograms, bar charts, scatter plots, and more using Matplotlib. Matplotlib has adaptability. It works nicely with many other Python libraries, including NumPy and pandas, and offers a wide variety of graphs and plots (datacamp, 2023).

- Seaborn: A Python module for data visualization is called Seaborn. High-level plotting capabilities are available to create visually appealing and educational visuals. It can be integrated with NumPy and Scikit-learn and is best suited for working with pandas' data structures (AlmaBetter, 2023).

- Scikit-learn: Python has a machine-learning package called Scikit-learn. It offers resources for selecting models, prepping data, supervised and unsupervised learning, and other tasks. It is made to be reliable, effective, and simple to use (AlmaBetter, 2023).

3.7.3. Tools

- Kaggle: It is an online community of machine learning engineers and data scientists run by a Google subsidiary. Users may identify datasets on Kaggle that they wish to utilize to create AI models, post datasets, collaborate with other machine learning engineers and data scientists, and participate contests to find solutions to data science problems (LLc, 2022).

3.8. Achieved results (screenshots of the application/screenshots of the results attained)

Importing Necessary Libraries

```
#Importing the Depedencies
import numpy as np
import pandas as pd
import plotly.offline as py
py.init_notebook_mode(connected=True)
import plotly.graph_objs as go
import plotly.express as px
from sklearn.preprocessing import StandardScaler
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats

#Sklearn
import sklearn
from sklearn.svm import SVC
from sklearn import metrics
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
#naive bayes
from sklearn.naive_bayes import GaussianNB
from sklearn.naive_bayes import MultinomialNB
from sklearn.naive_bayes import BernoulliNB
import warnings
warnings.filterwarnings('ignore')
```

Figure 16: Importing necessary libraries

Importing Datasets

LOADING THE DATA

```
# Read the CSV File Using Pandas read_csv function
Breast_Cancer_df = pd.read_csv('../input/breast-cancer-wisconsin-data/data.csv')
```

```
Breast_Cancer_df
```

Figure 17: importing dataset

Data Pre-processing

DATA CLEANING

```
Breast_Cancer_df.info()
```

```
Breast_Cancer_df.isna().sum()
```

```
Breast_Cancer_df.hist(figsize = (20, 20))
plt.show()
```

+ Code + Markdown

```
fig= px.histogram(Breast_Cancer_df, x='diagnosis',color='diagnosis', barmode='group')
fig.show()
```

Figure 18: data cleaning

Visualization

EXPLORATORY DATA ANALYSIS (EDA)

BASIC STATISTICAL DETAILS

```
Breast_Cancer_df.hist(figsize = (20, 20))
plt.show()
```

+ Code + Markdown

```
fig= px.histogram(Breast_Cancer_df, x='diagnosis',color='diagnosis', barmode='group')
fig.show()
```

```
fig = px.pie(Breast_Cancer_df, values='radius_mean', names='diagnosis', title='Pie chart to show to relation of malignant and be')
fig.show()
```

Pie chart to show to relation of malignant and benign

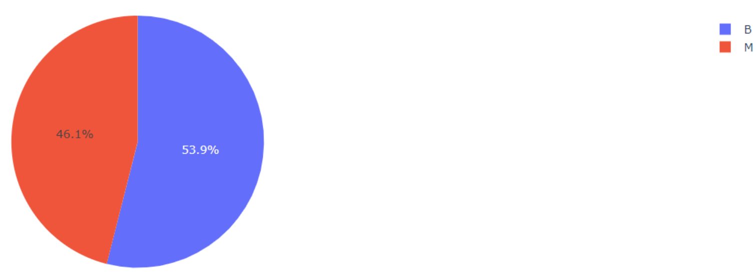

Figure 19: visualization

DATA MANIPULATION

```
diagnosis_coder = {'M':1, 'B':0}
Breast_Cancer_df.diagnosis = Breast_Cancer_df.diagnosis.map(diagnosis_coder)
Breast_Cancer_df
```

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean
0	1	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010
1	1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690
2	1	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740
3	1	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140
4	1	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800
...	...	...	...	...	...	...	...	...
564	1	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390
565	1	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400
566	1	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251
567	1	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140
568	0	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000

569 rows × 31 columns

Figure 20: data manipulation

```
fig, ax = plt.subplots(figsize=(15,15))
sns.heatmap(Breast_Cancer_df.corr(),ax=ax,annot=True,linewidth=.5)
plt.show()
```

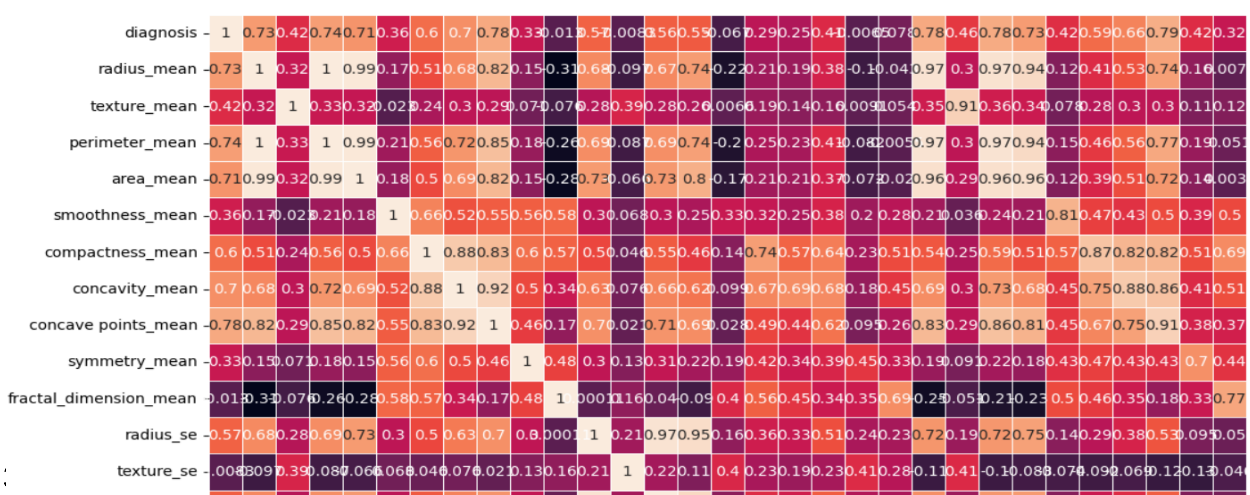

Figure 21: Heatmap to show correlation

Separating Train and Test Values

SEPERATING FEATURES AND TARGETS

```
X= Breast_Cancer_df.drop(columns='diagnosis', axis=1)
Y= Breast_Cancer_df['diagnosis']
```

```
X
```

Figure 22: separating features

Training and testing the models

SPLITTING DATASET INTO TRAIN TEST MODEL

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.3, random_state = 12)
```

```
print(X.shape, X_train.shape, X_test.shape)
```

```
(569, 30) (398, 30) (171, 30)
```

Figure 23: splitting train test data

LogisticRegression

LOGISTIC MODEL

```
model = LogisticRegression()
```

TRAINING THE LOGISTIC REGRESIION USING TRAINING MODEL

```
model.fit(X_train, Y_train)
```

```
▾ LogisticRegression
LogisticRegression()
```

```
# classification report
print(classification_report(Y_test, y_prediction))
```

```
              precision    recall  f1-score   support

           0       0.92      0.99      0.95       107
           1       0.98      0.86      0.92        64

    accuracy                           0.94       171
   macro avg       0.95      0.93      0.94       171
weighted avg       0.94      0.94      0.94       171
```

```
# Generate the classification report
report = classification_report(Y_test, y_prediction, output_dict=True)

# Convert the classification report dictionary to a DataFrame
report_df = pd.DataFrame(report).transpose()

# Plot the heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(report_df.iloc[:-1, :-1], annot=True, cmap="Blues", fmt=".2f", cbar=False)
plt.title('Classification Report Heatmap')
plt.show()
```

Figure 24: Logistic regression

KNearest Neighbours

K NEIGHBOURS CLASSIFIERS(KNN)

```
#fitting knn in training set
knn = KNeighborsClassifier()
knn.fit(X_train, Y_train)
```

```
▾ KNeighborsClassifier
KNeighborsClassifier()
```

```
#predicting result using testing data
Y_pred = knn.predict(X_test)
Y_pred
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1,
       0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1,
       1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0,
       0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1,
       0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0])
```

Figure 25: KNN

DECISION TREE

```
#fitting training data to model
dt_model = DecisionTreeClassifier(criterion='entropy', random_state = 0)
dt_model.fit(X_train, Y_train)
```

```
DecisionTreeClassifier
DecisionTreeClassifier(criterion='entropy', random_state=0)
```

```
#predicting result using testing data
Y_DT_pred = dt_model.predict(X_test)
Y_DT_pred
```

```
array([1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0,
       0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0,
       0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1,
       1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0,
       0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1,
       0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0])
```

Figure 26: Deciison trees

Naïve Bayes

NAIVE BAYES

```
nav_model = GaussianNB()
```

```
#fitting training data to model
nav_model.fit(X_train, Y_train)
```

```
▾ GaussianNB
GaussianNB()
```

```
#predicting result using testing data
nav_pred = nav_model.predict(X_test)
nav_pred
```

```
array([1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0,
       0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0,
       0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0,
       0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1,
       1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0,
       0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1,
       0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0])
```

Figure 27:NAive Bayes

Support Vector Classifiers

```
#fitting SVC to training set
svc = SVC()
svc.fit(X_train, Y_train)
```

```
▾ SVC
SVC()
```

```
#predicting result using testing data
Y_svc = svc.predict(X_test)
Y_svc
```

Figure 28: SVC

4. Conclusion

4.1. Analysis of the work done

In conclusion, the second coursework on the topic “Breast Cancer Classification” for the module artificial intelligence is done. he relevant topic and the method used were thoroughly and carefully researched. Researching on the selected topic “Breast Cancer Classification” throughout numerous books, journal articles, websites that are referenced. The analysis of the effort made is in developing the system that uses artificial intelligence, data science, and machine learning approaches to solve a major health care problem. This research course emphasizes how machine learning might revolutionize medical diagnosis in the real world.

The Breast Cancer Classification project is significant for its extensive evaluation, which covers important topics such data preparation, model selection, and evaluation. The main objective is to create a useful model that can classify early cancer detection, decreasing the risk of cancer that has moved on to an advanced stage. Model selection is followed by careful evaluation to ensure that the algorithms (Decision Trees, Naïve Bayes, Support Vector, and Logistic Regression) selected are appropriate for the goals of the project.

Important phases in the structured technique are thorough data exploration and understanding, data mining, data preparation, and model training as well as evaluation. The project's achievement is a result of each phase's contribution, which highlights the significance of a methodical and well researched strategy when applying machine learning to medical diagnostics in the real world.

4.2. Solution addresses real world problem

By focusing early cancer identification, the implementation of a machine learning-based Breast Cancer Classification system addresses a significant real-world issue. As breast cancer frequently remains undetected in its early stages, early identification is essential to improving patient survival rates. The system may detect small patterns symptomatic of cancer by using machine learning algorithms to examine a variety of patient records and imaging data. This allows for early treatment and prevention. This preventive strategy helps to address the growing global health issue caused by the rising number of occurrences of breast cancer, while also improving the results for individual patients.

Moreover, an approach based on machine learning performs a crucial role in preventing challenges related to advanced stages of breast cancer. The model reduces the possibility of complications and reduces the impact of the disease by empowering medical providers to detect and treat tumors early. The system also offers improved resources to medical practitioners for risk evaluation and treatment for patients. Its precision and methodical approach provide an affordable technique to enhance the diagnosis and prognosis of breast cancer, clearing up possibilities for more accessible and effective healthcare solutions globally. A few real-world problems that the model addresses are as follows:

- Early tumour treatment is feasible with its support.
- It will help expand the field of breast cancer research.
- It will assist in preventing any breast cancer problems.
- It will provide healthcare professionals with enhanced resources for patient care and risk evaluations.
- It will offer systematic, accurate, and useful data and It is economical.

4.3. Further Work

The specific application will be created for future work. We might eventually include the individual risk evaluations and grow the dataset. It is simple to examine with a mobile device. It is possible to apply deep learning to improve performance. It may be widely used for risk assessment studies in the healthcare industries. The model encourages a continuous learning environment by being updated with new patient data, which maintains the anticipated outcomes fresh and correct even when patient demographics and medical knowledge shift. It is possible to use real-time health data for ongoing risk updates.

Future developments in healthcare and technology might benefit greatly from the use of machine learning to the classification of breast cancer. Molecular imaging and three-dimensional mammography are examples of advanced imaging techniques that may be used to give richer, more comprehensive data for better model performance. Furthermore, the accuracy of predictive methods may be improved by using multi-modal data fusion, which combines data from many sources like genetics and patient history. It is possible to identify complicated patterns in complex datasets by using deep learning architectures that can automatically extract features and learn representations.

Additionally, the development of model interpretation will improve medical acceptability and make decision-making processes easier for people to understand. In order to resolve privacy issues, assure the ethical use of data, and provide strong validation frameworks, collaborations between researchers, physicians, and technological specialists are important. Machine learning models should be updated often as research into the biology of breast cancer grows, providing more accurate and customized diagnostic tools for breast cancer treatment.

References

AlmaBetter, 2023. *Python Libraries.* [Online]
Available at: https://www.almabetter.com/bytes/tutorials/python/popular-python-libraries
[Accessed 13 01 2024].

Arooj, S., 2022. Breast Cancer Detection and Classification Empowered with Transfer Learning. *Natinal Library of Medicine,* 10(II), p. 21.

Baum, H., 2012. *AN INTRODUCTION TO ARTIFICIAL INTELLIGENCE.* 2 ed. Ohio: The University of Cincinnati..

Brooks, R., 2023. *Reinforcement Learning.* [Online]
Available at: https://online.york.ac.uk/what-is-reinforcement-learning/#:~:text=Reinforcement%20learning%20(RL)%20is%20a,using%20feedback%20from%20its%20actions.
[Accessed 14 01 2024].

Codeacedemy, 2023. *Introduction to Pandas and Numpy.* [Online]
Available at: https://www.codecademy.com/article/introduction-to-numpy-and-pandas
[Accessed 13 01 2024].

Coleman, 2024. *What is 'Pseudocode',* India: The Economic Times.

datacamp, 2023. *Introduction to Plotting with Matplotlib in Python.* [Online]
Available at: https://www.datacamp.com/tutorial/matplotlib-tutorial-python
[Accessed 13 01 2024].

Dey, A., 2021. *Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms.* Pakistan, International Conference of AI.

Faud, W. M., 2018. *Early Detection of Breast Cancer Using Machine Learning.* 58 ed. Bangladesh: BRAC University.

Godfrey, 2023. Harnessing the power of AI to improve outcomes for people with breast cancer. *Demonstrating how AI can improve the accuracy of diagnosis*, 30 October, p. 8.

IBM, 2023. *Arificial Intelligence.* [Online]
Available at: https://www.ibm.com/topics/artificial-intelligence
[Accessed 14 01 2024].

Khandezamin, Z., 2020. Detection and classification of breast cancer using logistic regression. *Journal of Biomedical Informatics,* 111(10391), p. 55.

LLc, G., 2022. *Kaggle.* [Online]
Available at: https://en.wikipedia.org/wiki/Kaggle
[Accessed 13 01 2024].

LucidChart, 2024. *Flowchart.* [Online]
Available at: https://www.lucidchart.com/pages/what-is-a-flowchart-

tutorial#:~:text=A%20flowchart%20is%20a%20diagram,easy%2Dto%2Dunderstand%20diagrams.
[Accessed 10 01 2024].

Rahul Kumar Yadav, P. S. P. K., 2023. *Diagnosis of Breast Cancer using Machine Learning Techniques -A Survey.* 218 ed. India: International Conference on Machine Learning and Data Engineering.

Rajamani, R., 2020. *Breast Cancer Detection Using Logistic Regression.* [Online]
Available at: https://medium.com/swlh/breast-cancer-detection-using-logistic-regression-7bef0110aff5
[Accessed 15 01 2024].

Rane, N., 2020. Breast Cancer Classification and Prediction using Machine Learning. *IJERT,* 09(02), p. 580.

S., T., 2023. 3. A Breast Cancer Image Classification Algorithm with 2c Multiclass Support Vector. *Journal of Healthcare Engineering,* 22(1), p. 12.

Saha, D., 2023. A Brief Introduction to Artificial Intelligence. In: R. University, ed. *What is AI and how is it going to shape the future.* India: Ryerson University, pp. 1-9.

Team, I., 2023. *Algorithm.* [Online]
Available at: https://www.indicative.com/resource/programming-algorithm/#:~:text=Programming%20Algorithm%20Defined,exactly%20that%20way%20every%20time.
[Accessed 11 01 2024].

teradata, 2024. *Python.* [Online]
Available at: https://www.teradata.com/glossary/what-is-python#:~:text=Python%20is%20an%20interpreted%2C%20object,British%20comedy%20group%20Monty%20Python.
[Accessed 13 01 2024].

Topol, E., 2019. *Deep Medicine.* 1st ed. LA: La Jolla.

TutorialsPoint, 2024. *Artificial Intelligence.* [Online]
Available at: https://www.tutorialspoint.com/artificial_intelligence/index.htm
[Accessed 13 01 2023].

V, v., 2023. Breast Cancer Classification USing Logisticc Regresiion. *Logistic Regression,* 1006(2023), p. 6.

Xiao Jia, 1. S. X. Z., 2022. Breast Cancer Identification Using Machine Learning. *Theory and Application of Swarm Intelligence and Machine Learning,* 2022(8122895), p. 5.