**JAYPYEE INSTITUE OF INFORMATION AND TECHNOLOGY**

****

MINOR PROJECT – 1

ACADEMIC YEAR: - 2023-24

TITLE – BREAST CANCER CLASSIFICATION USING MACHINE

AND

DEEP LEARNING

BY: -

|  |  |
| --- | --- |
| NAME | ENROLLMENT NUMBER |
| 1. AYUSH AGRAWAL | 21104004 |
| 1. ABHILASH TYAGI | 21104019 |
| 1. OJASVI PANDEY | 21104027 |

SUPERVISOR: DR. ASMITA YADAV

TABLES OF CONTENTS

1. **General Introduction.**
2. **Problem Statement.**
3. **Motivation.**
4. **Literature Survey.**
5. **Layout of project.**
6. **Dataset.**
7. **Implement a model.**
8. **Future scope.**
9. **References.**

**ACKNOWLEDGEMENT**

I would like to place on record my deep sense of gratitude to **Dr. Asmita Yadav**, Assistant Professor (Senior Grade) in the department of Computer Sciences at Jaypee Institute of Information Technology for her generous guidance, help and useful suggestions.

I express my sincere gratitude to **Dr. Asmita Yadav**, Dept. Of Computer Sciences at Jaypee Institute of Information Technology, for her stimulating guidance, continuous encouragement and supervision throughout the course of present work.

**Name:** Ayush Agarwal

**Enrollment No.:** 21104004

**Name:** Abhilash Tyagi

**Enrollment No.:** 21104019

**Name:** Ojasvi Pandey

**Enrollment No.:** 21104027

* 1. **General Introduction-**

1) Breast cancer is a type of cancer that starts in the breast. Cancer starts when cells begin to grow out of control. Breast cancer cells usually form a tumour that can often be seen on an x-ray or felt as a lump. Breast cancer can spread when the cancer cells get into the blood or lymph system and are carried to other parts of the body.

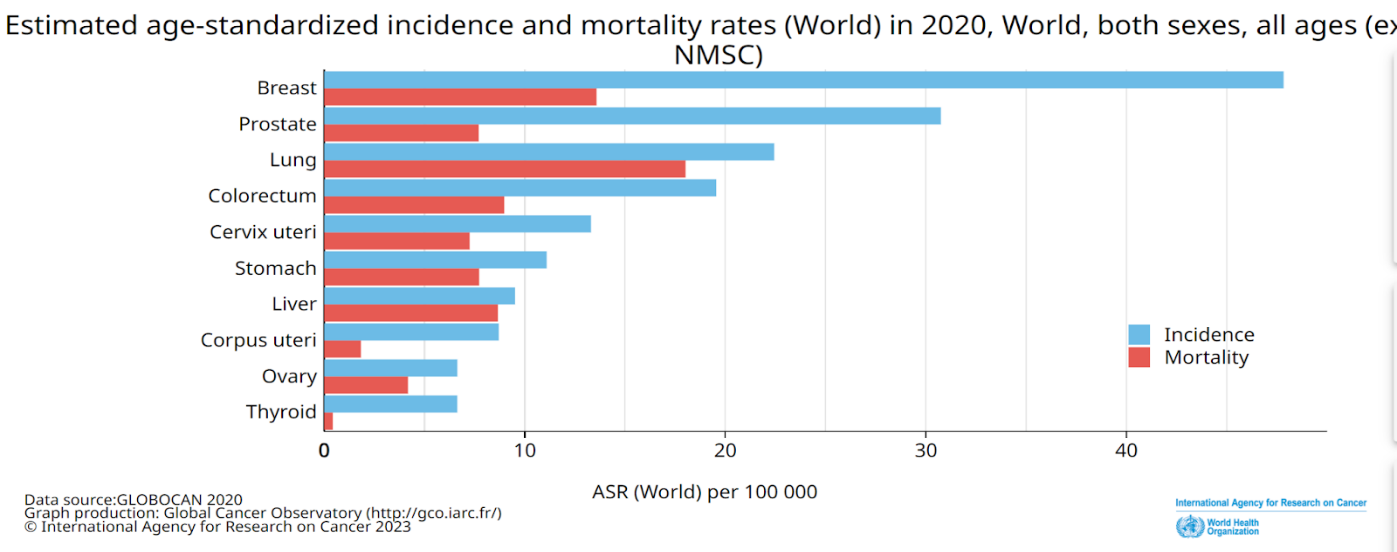
2) Early detection of breast cancer is pivotal in improving treatment outcomes and increasing the chances of survival for affected individuals. Conventional diagnostic methods, though effective, may have limitations in terms of sensitivity and efficiency. This underscores the need for advanced technological approaches to enhance the accuracy and efficiency of breast cancer detection.

3) Researching on how ML techniques are revolutionizing breast cancer detection and management, providing more effective and precise solutions for patients and healthcare professionals alike.

* 1. **Problem Statement**

With a staggering 93% chance of successful treatment when detected early, breast cancer remains a critical focus in healthcare. However, timely, and accurate diagnosis remains a challenge, this project aims to leverage these cutting -edge techniques to significantly improve the speed and precision of disease diagnosis, particularly in regions with limited access to specialized medical resources. Through this research, we strive to contribute to a more efficient and effective healthcare system, ultimately leading to enhanced patient outcomes

**1.3**   **Motivation behind the project**

  1) Breast cancer is the **most common occurring cancer** in the world according to global cancer observatory data. 

**2)** **Early Detection:** Accurate classification helps in the early detection of breast cancer. Detecting cancer at an early stage significantly improves the chances of successful treatment and a better prognosis

**3) Reducing Unnecessary Procedures**: Accurate classification reduces unnecessary biopsies and surgeries for benign tumours. Many breasts tumour is not cancerous, and misclassifying them malignant could subject patient to unnecessary stress, medical procedure and costs.

**4)Survival Rates**: Accurate classification helps in the early detection of breast cancer. Detecting cancer at an early at an early significantly improves the chances of successful treatment and a better prognosis.

**LITERATURE SURVEY**

**COLLECTIVE RESEARCH PAPERS SUMMARY: -**

These papers collectively focus on breast cancer detection and healthcare decision-making using advanced computational methods. They stress early detection's crucial role, employing techniques like deep learning, machine learning, and genetic algorithms. Machine learning algorithms, including SVM, KNN, and logistic regression, demonstrate high accuracy in breast cancer prediction. Feature selection and ensemble methods are highlighted for performance improvement. Deep learning, especially CNNs, proves effective in medical imaging. Overall, these studies underscore machine learning and deep learning’s transformative impact on healthcare decision-making, offering swift and accurate analyses for early diagnosis and treatment. They highlight diverse approaches and algorithms, applied in disease prediction, imaging, biomedicine, and drug discovery. These findings collectively affirm the vital role of machine learning in advancing breast cancer diagnosis and healthcare decisions.

**1)Significance of machine learning in healthcare: features, pillars and applications. [add reference no]**

Machine learning (ML) is revolutionizing healthcare by enabling data-driven decision-making, automation, and efficiency. This research paper explores the significance of ML in healthcare, highlighting its key features, pillars, and applications. ML's capacity to process vast medical data aids in disease diagnosis, personalized treatment, and predictive analytics. Pillars like data quality, algorithm development, and interoperability ensure its success. ML finds diverse applications, including drug discovery, healthcare management, telemedicine, and disease detection. Overall, ML enhances healthcare's accuracy, efficiency, and patient-centricity, promising improved outcomes in a rapidly evolving healthcare landscape.

2)Involvement of machine learning tools in healthcare decision making**. [add reference no]**

This review underscores the pivotal role of machine learning in swift and accurate healthcare decision-making, spanning disease prediction, medical imaging, biomedicine, and drug discovery. It highlights the continuous enhancement of algorithms, with deep learning excelling in complex feature extraction from medical images. The paper affirms machine learning's indispensability in modern healthcare and anticipates further revolutionization through ongoing advancements in scalable algorithms and artificial intelligence.

**3) JETIR: Breast cancer prediction using machine learning. [add reference no]**

The paper underscores the critical need for improved breast cancer detection beyond conventional screening methods. In essence, this research paper provides a holistic exploration of machine learning's applicability in breast cancer prediction. It accentuates the significance of data preprocessing, feature selection, scaling, model evaluation, dimensionality reduction techniques and interpretability for the development of robust and accurate prediction models.

**4)Breast cancer classification using machine learning: a comparative study. [add reference no]**

This research paper conducts a thorough comparative study on the application of machine learning techniques for breast cancer classification, shedding light on the effectiveness of various algorithms and methodologies in diagnosing this critical medical condition. It systematically evaluates the performance of multiple machines learning algorithms, including decision trees, support vector machines, k-nearest neighbours, logistic regression, and random forests based on its effectiveness and efficiency. Key aspects scrutinized in the comparative study include classification accuracy, sensitivity, specificity, and roc, providing a comprehensive overview of each algorithm's strengths and limitations.

**5)Classification Prediction of Breast Cancer Based on Machine Learning. . [add reference no]**

This research paper investigates breast cancer prediction using machine learning, emphasizing recall as a vital metric. It tests various models with different data splits, such as XGBoost, Random Forest, Logistic Regression, and KNN. XGBoost excels in an 8:2 split, while logistic regression performs best in a 7:3 split. Comparisons with existing research show XGBoost's superiority in recall and accuracy. The paper underscores its significance for breast cancer diagnosis improvement and suggests future exploration of deep learning algorithms for image data classification.

**6)Diagnosis of Breast Cancer using Machine Learning Techniques - Survey. [add reference no]**

This research paper offers a comprehensive exploration of machine learning and deep learning techniques for breast cancer diagnosis, emphasizing the significance of feature extraction, performance metrics, and dataset size. It identifies SVM and ensemble methods as robust traditional ML approaches and highlights CNNs' effectiveness in image-based diagnosis. Ultimately, the research aims to advance early breast cancer detection through advanced computational methods, making it a valuable contribution to the field.

**7)Breast cancer diagnosis in an early stage using** novel **deep learning with hybrid optimization technique. . [add reference no]**

This research paper addresses the critical issue of breast cancer detection from MRI images by introducing the novel approach, BPBRW with HKH-ABO, emphasizing early detection's importance. It reviews existing methods, introduces the proposed deep learning-based methodology, and highlights superior performance in classifying breast cancer types. While showcasing potential for early detection, the paper concludes with the need for further optimization in reducing design time and suggests future work in enhancing security mechanisms and clinical applications.

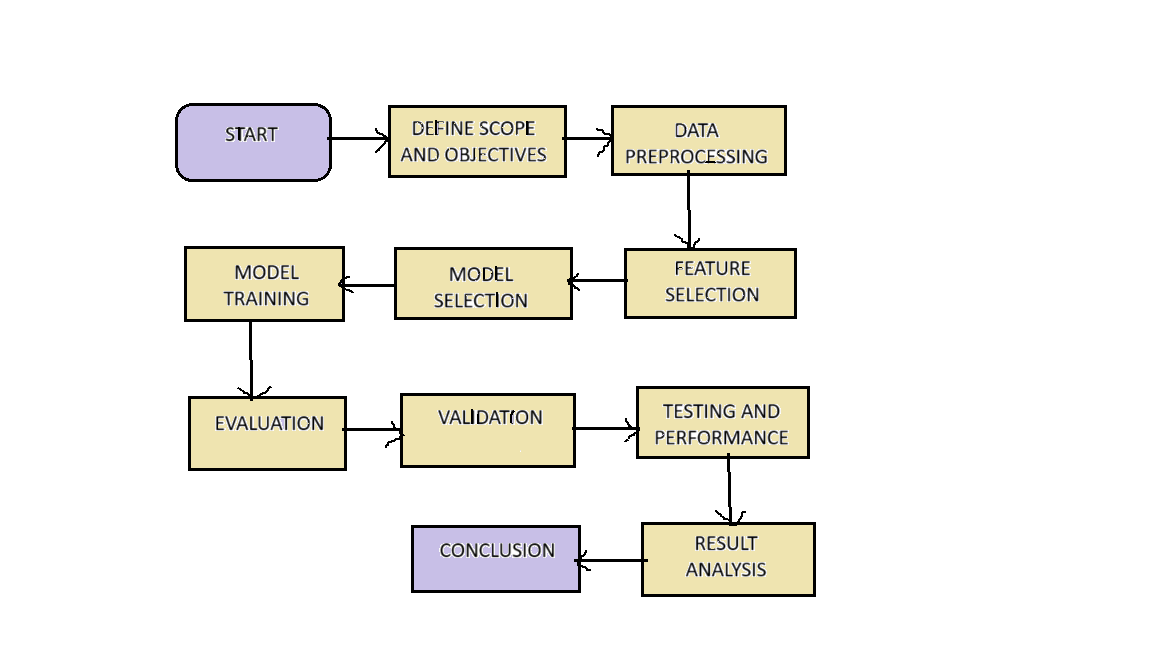
**8) Review paper on breast cancer detection using deep learning. [add reference no]**

This paper underscores the importance of early breast cancer detection, a critical global health concern for women. While traditional methods yield vital images, deep learning, specifically CNNs, emerges as a potent tool for accurate classification. Recent advances in CNN models demonstrate remarkable performance, reaffirming deep learning's superiority in breast cancer classification and its potential to enhance screening and diagnosis.

 9) Breast cancer classification and detection using machine learning**. [add reference no]**

This paper highlights the critical global need for early breast cancer detection, particularly in developing nations facing a rise in late diagnoses. It explores diverse techniques like ensemble methods, genetic algorithms, and blood analysis to improve prediction accuracy. Among machine learning models, KNN surpasses CART, random forest, and boosted trees on UCI data. The proposed system evaluates multiple models on the WDBC dataset, aiming to efficiently detect benign/malignant cancer from digitized biopsy images, with potential benefits for healthcare systems.

LAYOUT OF OUR PROJECT:

****

**IMPLEMENTATION OF MODEL**

 Our Project is designed to be implemented in **three phases** where -

* **Phase 1** includes breast cancer classification using three ML techniques namely KNN, Random Forests and Logistic Regression.
* **Phase 2** of our project aims at building our own neural network model by using tensorflow and keras in python.
* **Phase 3** Involves integration of the ideas of energy consumption into the domain of health care.
* **PHASE 1 - CLASSIFICATION USING ML TECHNIQUES**

**1)RANDOM FOREST ALGORITHM: -**

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

**Implementation techniques of random forest**

1. **Data Preparation**: Starting with a dataset containing features and target values.

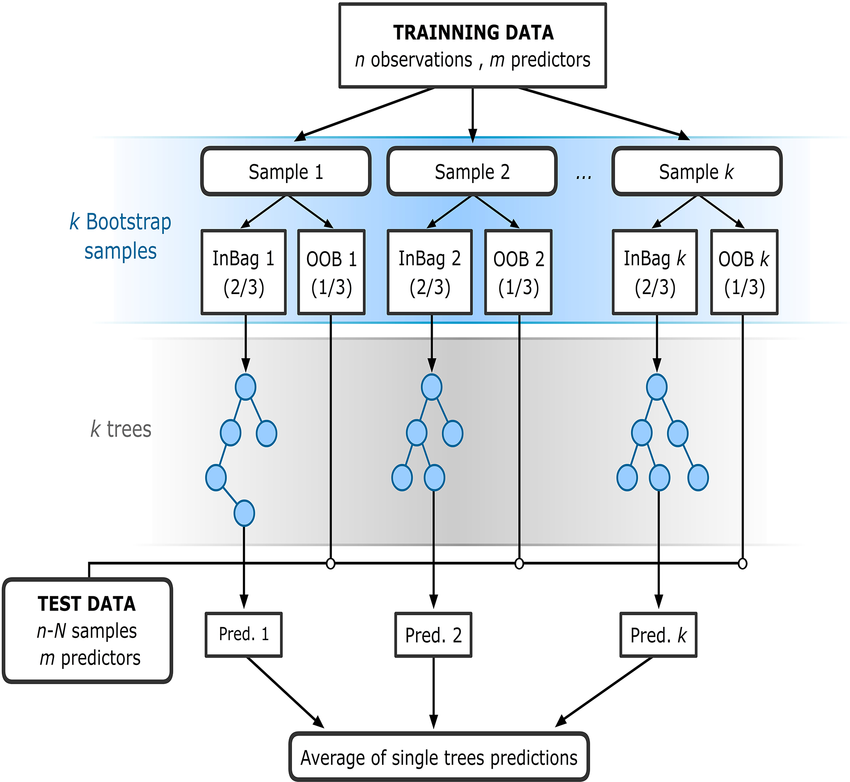
2. **Bootstrapping**: Randomly selecting the subsets of the data (with replacement) to create multiple training datasets.

3. **Feature Randomization:** At each tree node, we will select a random subset of features for split candidates.

4. **Build Decision Trees**: Creating a decision tree for each training dataset, using the selected features.

5. **Aggregate Predictions:** Combining the tree predictions for classification (majority vote) or regression (average).

6. **Output:** Obtaining the final ensemble prediction.



**2) KNN TECHNIQUE**

K-Nearest Neighbours (KNN) is a simple yet effective machine learning algorithm used for classification and regression tasks. It works based on the principle of finding the 'k' training examples in the feature space that are closest to the new data point and using them to predict the class or value of the new data point.

**Implementation Steps**:

1. Data Preprocessing including cleaning and preprocessing the dataset

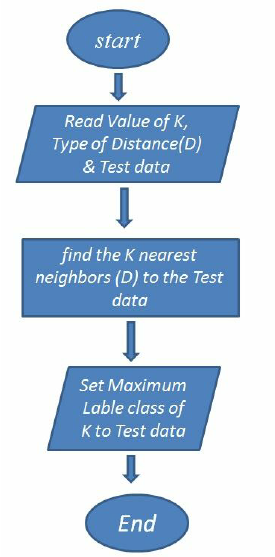
2. Splitting the data and dividing dataset into training and testing set

3. Implementing the KNN Algorithm and calculating Euclidean distance and k nearest neighbours.

4. Classification for each new data point, classifying it based on the majority class among its   'k' nearest neighbours.

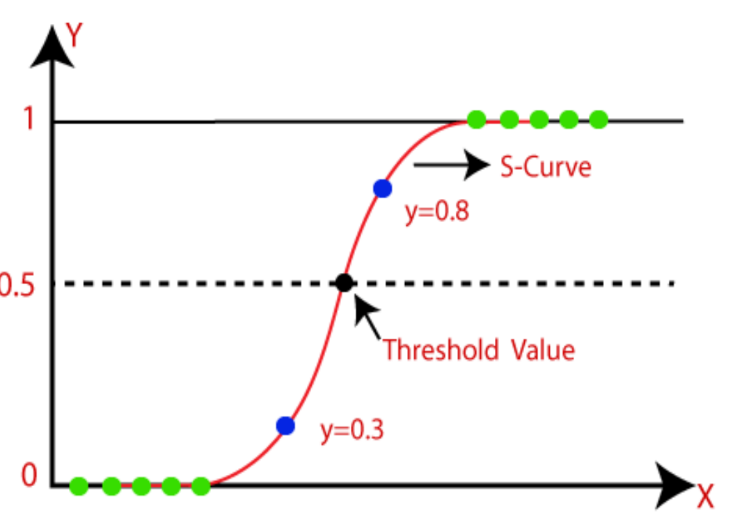
5.Evaluating the model and its performance and tuning its hyperparameters and now comparing it with other models

 flowchart



**3)LOGISTIC REGRESSION-:**

Logistic regression was introduced by statistician DR Cox in 1958 and so predates the field of machine learning. It is a supervised machine learning technique, employed in classification jobs (for predictions based on training data). Logistic Regression uses an equation like Linear Regression, but the outcome of logistic regression is a categorical variable whereas it is a value for other regression models. Binary outcomes can be predicted from the independent variables.

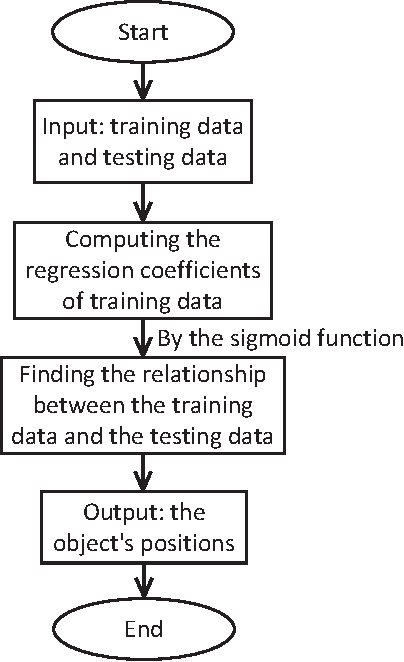
In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

**Steps in Logistic Regression:**

To implement the Logistic Regression using Python, we will use the following mentioned steps. Below are the steps: -

* The Data Preprocessing step involves cleaning of raw data and normalizing it .
* Build a logistic regression model by fitting a logistic curve to the training data, learning the optimal weights for each feature
* Predicting the test result using the trained model and applying threshold to the predicted probabilities to classify instances into one of the two classes.
* Test accuracy of the result (Creation of Confusion matrix) and evaluating its efficiency and effectiveness.
* Visualizing the test set result.

**WORKFLOW CHART-**



* **PHASE 2- BUILDING A NEURAL NETWORK TRAINING MODEL USING DEEP LEARNING BY TENSORFLOW AND KERAS.**

Considering the objective of building our own model for breast cancer detection, we are currently in the process of exploring the capabilities of TensorFlow and Keras in the realm of deep learning. This approach aims to harness the potential of neural networks to significantly improve the accuracy and effectiveness of our diagnostic system.

TensorFlow is an open-source machine learning library that facilitates the creation and training of various machine learning models, including neural networks.

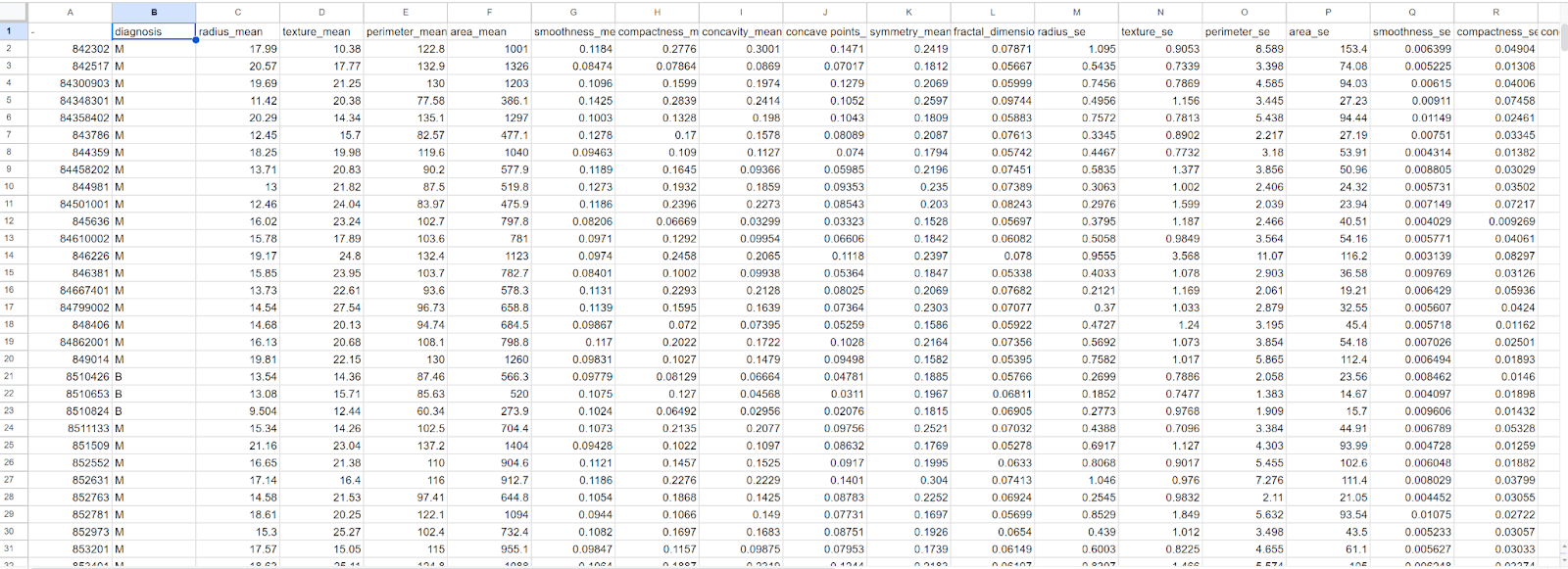
Keras is a high-level neural network which serves as an interface for TensorFlow and other backend libraries, enabling rapid prototyping and experimentation in deep learning projects

* **PHASE 3- ENERGY CONSUMPTION:**

 Accurate and efficient algorithms are crucial for optimizing energy consumption across various sectors. They enable better decision-making, reduce waste, and contribute to sustainability efforts by lowering energy usage and associated costs. As technology and data continue to advance, the role of algorithms in achieving energy efficiency will become increasingly significant

Subsequently, we will conduct a comparative analysis between our trained deep learning model and various conventional machine learning approaches that we have implemented in phase 1 of our project based on its effectiveness and efficiency to ascertain its impact on diagnostic decisions.

**DATASET-**



1)The above dataset is **Breast cancer Wisconsin diagnostic** data taken from UCI Machine Learning Repository.

2)Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. **They describe characteristics of the cell nuclei present in the image.**

3) **Data Size**: The dataset typically includes around **569 instances or sample**s.

4) **Data Features**: The dataset **contains 30 features**, including measures of cell nucleus characteristics such as radius, texture, smoothness, compactness, and concavity.

 5)**Target Variable**: The target variable is **binary**, with **'M'** representing **malignant** tumours and **'B'** representing **benign tumours**

6)    All feature values are recorded with **four significant digit**s.

* **Missing attribute value**s: none
* **Class distribution**: 357 benign, 212 malignant

7) The **important features selected** for our project are -:

**Concave Points Worst**:  Represents the severity of concave portions of the contour of the tumour in its worst condition. It measures the number of concave regions in the outline of the tumour and can be indicative of the irregularity of tumour shape.

**Area Wors**t: Refers to the area of the tumour in its worst condition. This feature quantifies the size of the largest lesion observed on the tumour, providing information about its extent.

**Area SE (Standard Error):** The standard error of the area measurement. This can indicate the consistency or reliability of the area measurement across different samples.

**Texture Worst**: Texture refers to the variation in grayscale intensity in the image of the tumour. "Texture Worst" represents the measurement of texture in the most malignant part of the tumour, reflecting the irregularity or variation in pixel intensities.

**Texture Mean**: Measures the texture of the tumour, but it represents the mean texture value across the entire tumour image.

**Smoothness Worst**: Quantifies the local variation in radius lengths in the worst portion of the tumour. High smoothness values indicate a more uniform tumour boundary.

**Smoothness Mean**: Represents the mean smoothness of the tumour boundary across the entire image.

**Radius Mea**n: Measures the mean of distances from the centre to points on the perimeter of the tumour. This parameter provides information about the size of the tumour.

**Symmetry Mean**: Quantifies the symmetry of the tumour. A higher symmetry value indicates a more symmetrical tumour shape, while lower values may suggest irregularities in shape.

**FUTURE SCOPE-**

The future scope for breast cancer classification using machine learning and deep learning holds significant promise and potential for further advancements such as: -

1. To achieve higher diagnostic accuracy, reduce the rate of false positives and false negatives in breast cancer detection, we can look forward to   implementing more fine-tuning models and incorporating more advanced algorithms such as genetic algorithms.

2.Focus on integrating ML-powered breast cancer classifiers into wearable devices or remote monitoring systems, thus enabling real-time monitoring of patients' health and treatment progress, leading to timely interventions.

**REFRENCES Write the refrence in the right Wy, for this check any paper reference section**

1.<https://www.researchgate.net/publication/361105025_Significance_of_machine_learning_in_healthcare_Features_pillars_and_applications>

**2**. https://pubmed.ncbi.nlm.nih.gov/33575021/

**3.** <https://www.jetir.org/papers/JETIR2211084.pdf>

4https://www.researchgate.net/publication/346617710\_Breast\_cancer\_classification\_using\_machine\_learning\_techniques\_a\_comparative\_study#:~:text=The%20used%20dataset%20is%20Wisconsin,98%25%20for%20breast%20cancer%20classificatio**.**

**5.** https://www.hindawi.com/journals/cin/2023/6530719/

6. <https://www.sciencedirect.com/science/article/pii/S1877050923001229#:~>:

text=This%20research%20investigates%20the%20detection,well%20as%

20mammography%20imaging%20datasets.

7. <https://link.springer.com/article/10.1007/s11042-022-12385-2>

8. <https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012071>

9 https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012071

10.<https://www.javatpoint.com/machine-learning-random-forest-algorithm>

**11.**https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

**12** https://www.javatpoint.com/logistic-regression-in-machine-learning

**13..** <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>