**Title: Breast Cancer Detection Using Machine Learning**

# Brief Idea

## Overview

Breast cancer remains one of the leading causes of cancer-related deaths among women worldwide, making early detection crucial for effective treatment and improved survival rates. Traditionally, screening methods such as mammograms have played a significant role in identifying breast cancer; however, these methods are not infallible and can result in false positives or negatives. The integration of machine learning (ML) into breast cancer detection presents a transformative opportunity to enhance diagnostic accuracy and efficiency.

## Application of Machine Learning

Machine learning leverages vast datasets to identify patterns that may go unnoticed by human observers. For instance, a study conducted by Nayak & Gope (2024) using the Breast Cancer Wisconsin Diagnostic dataset revealed that Support Vector Machine (SVM) algorithms achieved an impressive accuracy rate of 97.2%. This remarkable performance showcases the potential of machine learning techniques such as Random Forest, Logistic Regression, Decision Trees, and K-Nearest Neighbors to revolutionize breast cancer diagnostics.

## Scope of Research

The necessity for advanced detection methods is underscored by the increasing incidence of breast cancer globally. Integrating artificial intelligence with imaging technologies like mammograms and MRIs enhances radiologists' capabilities by identifying subtle abnormalities that may indicate early-stage cancers. By augmenting human expertise rather than replacing it, AI serves as a vital ally in improving patient outcomes through timely interventions.

## Problem Statement

Predicting human disease is challenging, as health issues depend on the predicted outcome. Breast cancer prediction is difficult due to varying symptoms. Machine learning can effectively detect breast cancer, but unseen data may cause misclassification. Meta-learning architecture may improve prediction accuracy by utilizing internal parameters of state-of-the-art models.

# Research Questions, Aim and Objectives

## Research Questions

The research questions farmed from the present research are as follows:

1. Can machine learning help detect breast cancer with higher accuracy?
2. How do analytical approaches help prepare data for the detection of breast cancer?
3. How does the user interface facilitate in detection of breast cancer?

## Aim

The aim of the project is to analyze the symptoms of patients for the early diagnosis of breast cancer by detecting it using machine learning.

## Objectives

The objectives of the research are stated below:

* Obtain information on the signs and risk factors linked to breast cancer, as well as analyze current studies to determine methods for predicting breast cancer.
* Gathering a breast cancer dataset containing symptoms and choosing cutting-edge machine learning models based on a thorough literature review.
* Develop a meta-learning model with state-of-the-art machine learning models, both in their default settings and optimized through cross-validation and hyperparameter tuning.
* To utilize advanced models and meta-learning for predicting breast cancer, comparing performance metrics, and studying how model tuning impacts detection accuracy and precision.
* To recognize the obstacles in enhancing research and the gaps in research when compared to current literature, and to compile the final project documentation.

# Project’s Integral Components

## Data

### Data Overview

The dataset for the present research to detect breast cancer will be taken from UCI (Mangasarian, 1992). The data contains the historical records of patients who have malignant breast cancer and those who were healthy. The data link is ***https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)***

### Features

The dataset contains 699 records of historical patients who have been taken into the testbench to capture symptoms. In this data, 9 symptoms have been recorded regarding breast cancer and the outcomes are stored in the “class” variable. The class variable contains two classes namely class 2 (Benign) and class 4 (Malignant).

### Data Management

#### Document Control

Github will be used to update the code and data there. In this context, the code will be sequentially updated on the following GitHub profile:

https://github.com/Amudalapallyparimala24/Msc-Data-science-project.git.

#### Security and Storage

The code will receive updates on Github two times per month. The information will be kept in that location and security measures will be managed by GitHub.

#### Data Ethics

The data will be sourced from UCI, a renowned platform known for providing a wide range of datasets. As UCI is a public data repository, the collection of data and reusing it from Kaggle will not raise any ethical issues.

#### Data Information and Attributes

The information of the data are as follows:

* Data Size: 84.5 KB
* Approximate Size of Code (In Jupyter Notebook Format 🡺 IPYNB): 3 MB
* Total Data Records: 699
* Format of Data: DATA format file
* Origin of the Data: UCI (University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg)

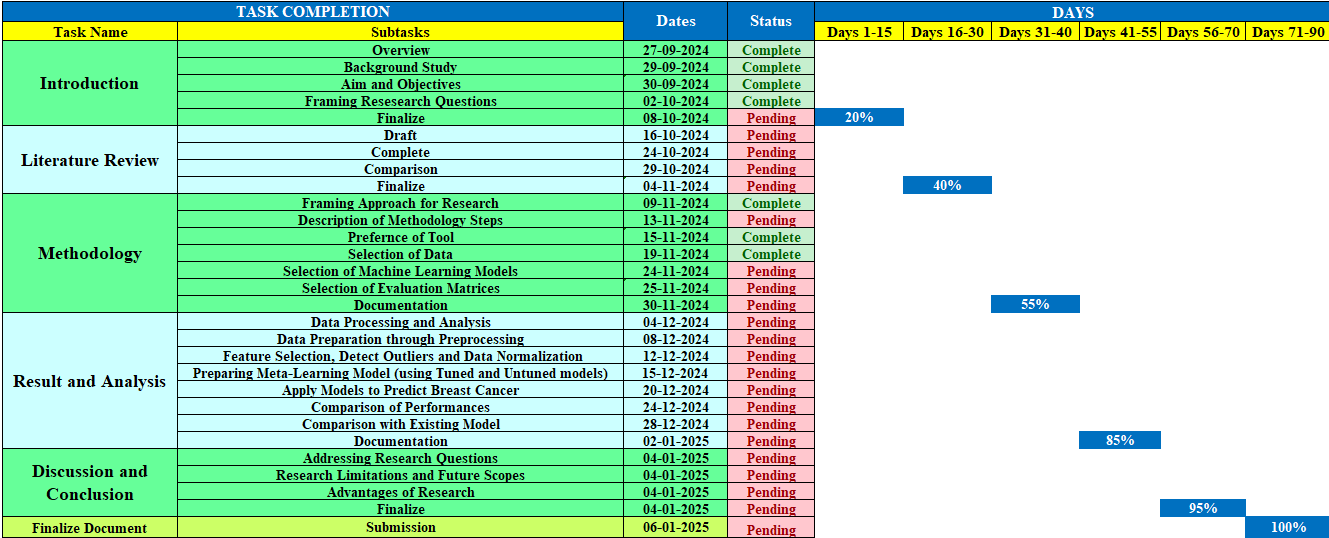
## Tool to be Used

Python will be utilized as the primary programming language in this research study to process the artefact for the purpose of accurately identifying potential signs of breast cancer by employing advanced machine learning algorithms.

# Planning

## Gantt Chart

The Gantt chart for the project is presented below with approximate completion dates:



## Explanation

The steps of the Gantt chart are explained below:

* The project will commence with preliminary studies, encompassing the identification of challenges, exploration of scopes, as well as the establishment of clear aims and objectives to guide further actions and decisions.
* Following this initial phase, the subsequent step involves a comprehensive examination of existing research, involving a meticulous critique to pinpoint areas of weakness or gaps that require further exploration and development.
* Advancing from the literature review, the project will progress to the formulation of a robust methodology, encompassing the detailed structuring of a framework, and careful selection of tools, algorithms, and evaluation techniques.
* Subsequently, leveraging the established methodology, the creation of the artefact will take place, aligning closely with the proposed framework to ensure coherence and effectiveness, followed by an in-depth analysis of the outcomes to derive meaningful insights that will be presented to stakeholders.
* Lastly, the project will culminate with a rigorous performance evaluation phase, where the achieved outcomes from the artefact will be scrutinized in tandem with addressing the core research questions, ensuring that the project's objectives have been met effectively.

# References

Agrawal, M. & Jain, V., 2022. Prediction of Breast Cancer based on Various Medical Symptoms Using Machine Learning Algorithms. 6th International Conference on Trends in Electronics and Informatics (ICOEI), pp. 1242-1245.

Anklesaria, S., Maheshwari, U., Lele, R. & Verma, P., 2022. Breast Cancer Prediction using Optimized Machine Learning Classifiers and Data Balancing Techniques. 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA, pp. 1-5.

Mangasarian, O., 1992. Breast Cancer Wisconsin (Original) Data Set. [Online]

Available at: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

Rovshenov, A. & Peker, S., 2022. Performance Comparison of Different Machine Learning Techniques for Early Prediction of Breast Cancer using Wisconsin Breast Cancer Dataset. 3rd International Informatics and Software Engineering Conference (IISEC), pp. 1-5.

Telsang, V. A. & Hegde, K., 2020. Breast Cancer Prediction Analysis using Machine Learning Algorithms. International Conference on Communication, Computing and Industry 4.0 (C2I4), pp. 1-6.

Nayak, S., & Gope, D. (2024). Machine Learning Algorithms For Breast Cancer Prediction, Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050921014629