PROJECT PROPOSAL

Precision-Driven Breast Cancer Diagnosis: Feature Engineering Meets Machine Learning

Team Members:

Likitha Magham

Jyothsna Allu

Pavana Manohari Gubbala

Dataset:

Breast Cancer Diagnosis Database :
https://www.kaggle.com/datasets/reihanenamdari/breast-cancer

Overview of the Project:

Breast cancer is the primary cause of cancer-related mortality in women, and early detection greatly improves treatment outcomes. This research aims to create and compare several machine learning models, such as ensemble methods and sophisticated algorithms, to predict the prognosis of breast cancer using clinical and diagnostic data. We want to uncover significant factors impacting patient outcomes by examining characteristics such as age, tumor size, regional node involvement, and survival duration.

Our goal is to improve predictive accuracy through feature engineering and machine learning approaches, hence offering significant insights to healthcare practitioners. This project will examine several machine learning models to discover the most effective method for assessing patient prognosis and survival probability, thereby assisting in data-driven decision-making for breast cancer treatment.

Reasons and Motivations:

Contributing to breast cancer research through data-driven insights offers a unique opportunity for us to enhance public health research and make a meaningful impact. We can gain valuable hands-on experience applying data science techniques to real-world problems by designing a

data mining pipeline. Our work aims to improve decision-making by providing actionable insights that medical professionals can use to manage patient care better, ultimately leading to more informed choices and better outcomes in breast cancer treatment.

Related Work:

Recent developments in deep learning (DL) and machine learning (ML) have greatly aided in the early identification and detection of cancer. The classification of breast cancer has been extensively studied using machine learning approaches, including Random Forest, Support Vector Machines, and Neural Networks. Research has demonstrated that feature selection and ensemble methods increase prediction accuracy. Litjens et al. (2017) highlighted the growing role of deep learning, particularly convolutional neural networks (CNNs), in medical image analysis, including the detection of breast cancer in mammograms. Similarly, Ragab et al. (2019) demonstrated how CNNs combined with support vector machines can significantly improve breast cancer detection accuracy. On the other hand, Mohapatra and Mohanty (2021) evaluated multiple machine learning techniques, showing that XGBoost achieved strong performance in predicting breast cancer from clinical data. Together, these studies emphasize the importance of leveraging advanced machine learning algorithms to enhance the accuracy and reliability of breast cancer diagnosis.

Data Plan:

The dataset contains 4,024 female breast cancer patients with clinical, demographic, and survival-related features like age, tumor size, stage, treatment, and survival time. Data preprocessing includes handling missing values, removing outliers, scaling numerical data, and encoding categorical variables. To handle class imbalance, **SMOTE** will be applied. These steps aim to improve the accuracy of statistical and machine learning models. The dataset comes from a public breast cancer repository, making it useful for real-world clinical research.

Implementation Plan:

We will start by analyzing feature distributions, correlations, and missing values to better understand the dataset and create visualizations to help interpret data. We will implement several machine learning models using training and testing sets, including two from the following: K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest, Logistic Regression, and Gradient Boosting Algorithms (XGBoost, LightGBM). To optimize performance, we will use hyperparameter tuning with GridSearchCV, RandomizedSearchCV, and k-fold cross-validation to minimize overfitting. Feature importance analysis will also be performed to identify key predictive factors. In the final stage, we will evaluate the models using accuracy, precision, recall, F1-score, and the confusion matrix to select the most effective approach.

Programming Environment:

The project will be developed using Python, incorporating several essential libraries. Scikit-learn will implement machine learning algorithms, Pandas will be used for data manipulation, and NumPy will assist with numerical computations. We will use Matplotlib and Seaborn for data visualization to communicate key insights. We will use the Imbalanced-learn library to manage potential class imbalances in the dataset and ensure our models are trained effectively, even with uneven class distributions.

Evaluation Plan:

The evaluation strategy will focus on measuring model performance using metrics, including accuracy, precision, recall, and F1-score, to assess the effectiveness of classification. The confusion matrix will analyze classification errors and identify where the models are misclassifying. Cross-validation (k-fold) will be implemented to reduce overfitting and ensure the models generalize well. Also, we will compare the performance of our models with a baseline model to quantify the improvements achieved through machine learning techniques.

Plan for Group Collaboration:

Our team will ensure communication and efficient project management for group collaboration. We will hold weekly virtual meetings to track progress, discuss challenges, and align on the next steps. Google Workspace will be used for collaborative writing and project tracking, allowing the team to stay organized and share documents in real-time. Code management will be handled through Git and GitHub for version control, ensuring synchronization of all team members. The implementation will primarily occur in Jupyter Notebook, which will be used for coding, data preprocessing, and model training.

Timeline:

Task No	Duties	Start date	Milestone
Task 1	Discussion and selection of the project topic	2/17/2025	2/20/2025
Task 2	Project proposal	2/21/2025	2/28/2025
Task 3	Dataset exploration and preprocessing	3/1/2025	3/7/2025
Task 4	Feature selection and engineering	3/8/2025	3/16/2025
Task 5	Model selection and validation	3/17/2025	3/22/2025

Task 6	Progress report	3/22/2025	3/24/2025
Task 7	Performance comparison and interpretation	3/24/2025	4/1/2025
Task 8	Visualization and Project Report	4/2/2025	4/8/2025
Task 9	Final Project Presentation	4/8/2025	4/18/2025

References:

- Litjens, G., et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60-88.
- Ragab, D. A., et al. (2019). Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ*, 7, e6201.
- Mohapatra, P., & Mohanty, S. (2021). Performance evaluation of machine learning techniques for breast cancer prediction using data analytics. *Journal of King Saud University - Computer and Information Sciences*, 33(6), 619-628.
- Esteva, A., Kuprel, B., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115-118.
- Chaurasia, V., & Pal, S. (2017). A novel approach for breast cancer detection using data mining techniques. *International Journal of Innovative Research in Computer and Communication Engineering, 5*(2), 241-246.