Machine Learning

Semester Project Proposal: A Comparative Study of Different for Breast Cancer Detection

Student Names	Laiba Binte Mazhar Mahnoor Athar Shanza Bakht
Registration Numbers	222403-MSDS-1 222391-MSDS-1 222594-MSDS-1
Class	MSDS-I
Teacher Name	Dr. Amaullah Yasin

Introduction:

Breast cancer is the second most frequent cancer in women worldwide after skin cancer. Both men and women can get breast cancer, although women are much more likely to do so. Breast cancer detection and treatment have advanced thanks to significant investment for research and awareness campaigns. With earlier identification, a novel customised approach to therapy, and a better knowledge of the illness, breast cancer survival rates have improved, and the number of fatalities linked to the disease is rapidly reducing. As a result, we made the decision to contrast and evaluate several machine learning approaches for the effective identification of Breast Cancer using Mammographic Masses.

Literature Review:

After surveying the literature available on this topic, the most common techniques used for breast cancer detection is Support Vector Machines (SVM), Naïve Bayes and Decision Tree. [1] used the WBCD dataset and implements Random Forest, Naïve Bayes, SVM and KNN algorithms. This is good to understand the basics, but this does not discuss any pre-processing or non-linearities of the data. While [2] using the same dataset, implements Multilayer Perceptron, KNN, Classification and Regression Trees (CART), Gaussian Naïve Bayes and SVM algorithms and concludes that Multilayer Perceptron is the best keeping in view the non-linearities of the data. The focus of [3] is on different classification techniques implementation for data mining in predicting malignant and benign breast cancer. This uses Ada Boost M1, Decision Table, J Rip, Lazy IBK, Logistics Regression, Multiclass Classifier, Multilayer Perceptron, Naive Bayes, Random forest and Random Tree are analyzed on WBCD.

Dataset:

A comparative study will be done on a dataset i.e., **Mammographic Mass Data Set** from UCI Machine Learning Repository [4]. Mammography is now the most efficient way to test for breast cancer. This dataset is used as a training set to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age.

This dataset has 961 instances and 6 attributes: BI-RADS assessment, age, shape, margin, density and severity. The class distribution is benign: 516 and malignant: 445. Whereas Missing Attribute Values are there: BI-RADS assessment has 2, Age has 5, Shape has 31, Margin has 48, Density has 76 and Severity has 0.

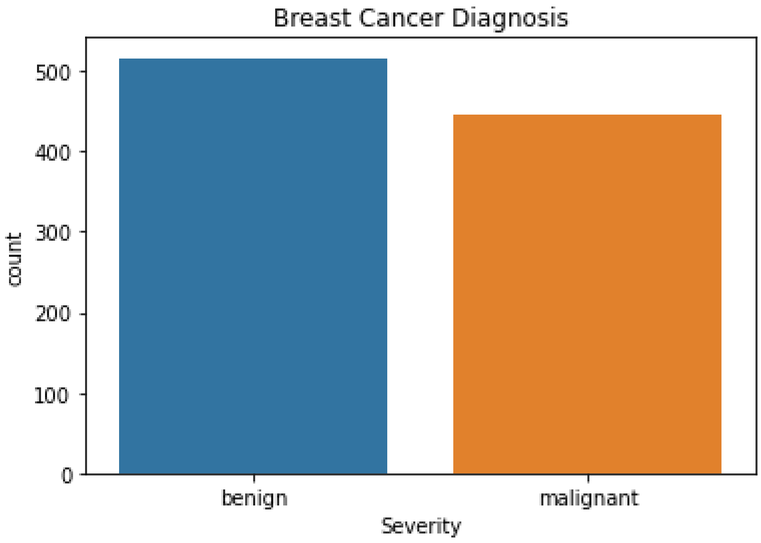

Figure 2: Class Distribution of Dataset

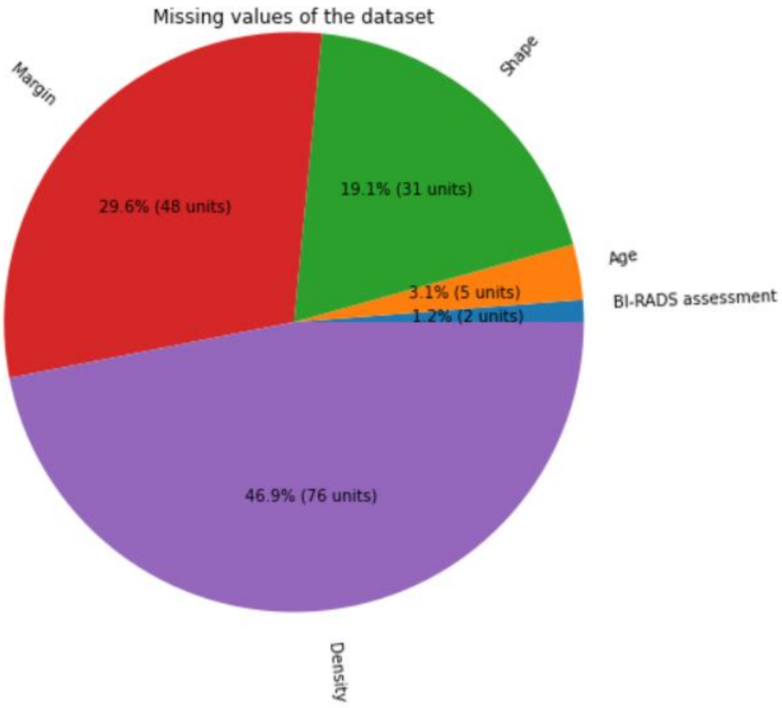

Figure 1: Missing Values of Dataset

Proposed Methodologies:

A survey will be done to relate the performances of various machine learning algorithm and predicts which technique is best for saving valuable life in terms of accuracy by the early detection of malignant tumour. A short-list of algorithms we will be utilizing are: Support Vector Machine (SVM), Random Forest, Logistic Regression, K-Nearest Neighbours (KNN), Naïve Bayes, Decision Tree and Adaptive Boosting (AdaBoost), Classification and Regression Trees (CART) and Multilayer Perceptron (MLP). Results obtained using the said techniques will be analysed based on different performance metrics.

References:

[1] *Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification*. (2018, December 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/8610632

[2] University of Toledo. (2019, June). *A Comparative Analysis of Nonlinear Machine Learning Algorithms for Breast Cancer Detection - Volume 9 Number 3 (Jun. 2019) - IJMLC*. http://www.ijmlc.org/index.php?m=content&c=index&a=show&catid=85&id=921

[3] Kumar, V. (2019, February 11). *Prediction of Malignant & Benign Breast Cancer: A Data Mining*. . . arXiv.org. https://arxiv.org/abs/1902.03825

[4] *UCI Machine Learning Repository: Mammographic Mass Data Set*. (n.d.). Retrieved October 31, 2022, from https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass