Harnessing the Power of Machine Learning for Breast Cancer Prediction

Nammiro Zaharah Bidin Reg: No: 21/U/1614 Std: No: 2100701614
Katwebaze Emmanuel Reg: No: 21/U/0615/PS Std: No: 2100706815
Nabbona Prossy Reg: No: 21/U/11450/PS Std: No: 2100711450
Mwegyesa Saul Reg: No: 21/U/0567 Std: No: 2100700567
Ssenteza Emmanuel Reg: No: 21/U/13955/PS Std: No: 2100713955

Abstract

Breast cancer is a serious health threat for women worldwide. Finding it early is key to successful treatment. Machine learning is a powerful tool that shows promise in detecting breast cancer early. This study compared five machine learning methods (SVM, Random Forest, Logistic Regression, XGBoost, and Neural Network Classification) as well as a deep learning method i.e. Convolutional Neural Network on the Breast Cancer Wisconsin (Diagnostic) dataset. We wanted to see which method worked best for predicting breast cancer. SVM came out on top with an impressive 98% accuracy, precision and recall.

I. Introduction

This report focuses on using machine learning to improve breast cancer diagnosis. Breast cancer is now the most common cancer in women. Cancer cases and deaths are rising worldwide, highlighting the need for better prevention and treatment. Information technology advancements are changing healthcare, especially cancer care. This study compares five learning algorithms to see which is most effective in predicting and diagnosing breast cancer. The goal is to find the best algorithm based on how well it identifies cancer, using different measurement techniques.

II. Proposed AI Methodology

A. Methodology

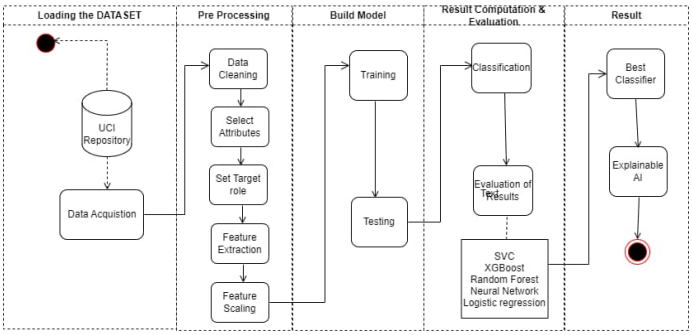

Fig. 1: Process Flow Diagram

We begin our work with obtaining the dataset from the UCI machine learning repository. We then carry out preprocessing which involves 5 steps which include, data cleaning, selecting the features, setting our target, feature extraction and feature scaling. We then move on to building the model which involves training and testing it. In this process, some hyper-parameter tuning takes place where we try to utilize the GridSearchCV algorithm to test and select various hyper-paramaters. We then compare and evaluate our results using the different selected models and finally obtain the best performing model from our classifier options. Our methodology ends with using an explainable AI technique to bring transparency to our model.

B. Dataset Description

The dataset, Breast Cancer Wisconsin (Diagnostic) Dataset, contains 569 records of various cell instances and 18 different features representing features of the cells inside the breast lump.

Factors considered when selecting the dataset

1) Size of the dataset: a good amount of data provides reliable predictions as it increases the domain of your training model.
2) Dimensionality: This is looked at in two ways: the vertical size of the dataset which represents the amount of data we have. The horizontal size which represents the number of features.
3) Context of the dataset: This involves basically picking a dataset that's in relation to the topic you are tackling.
4) Understandability: This involves selection of a dataset with features that can easily be understood and hence used to properly train the model you are creating.

C. Data Preparation and Exploratory Data Analysis

First we explored the dataset to discover the presence of any null values and it was determined that the dataset had zero null values.

We were able to visualize the diagnosis and determine the number of Benign or Malignant cases in the dataset.

Then to further prepare the data for use, two activities were performed; the diagnosis column which was originally in form of ”B” and ”M” was transformed into a column of booleans and secondly, the column was shifted from its position to the last column of the dataset.

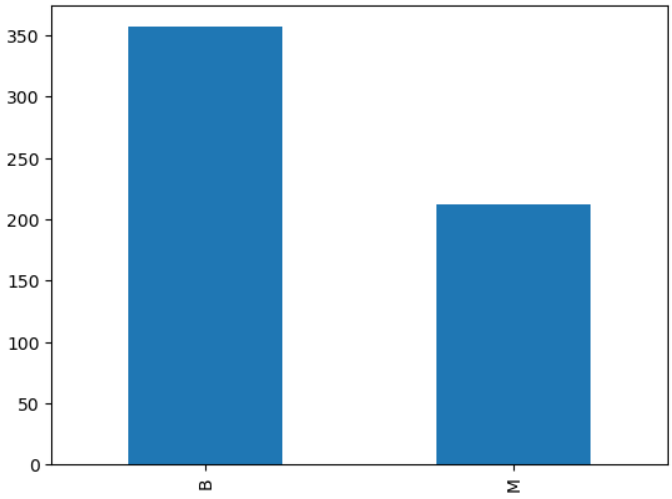

Fig. 2: Number of Cases

Note that the dataset was split into 80% training data and the remaining 20 percent was used for testing. Feature scaling was then performed on the training set to put all features on the same scale.

D. AI model selection and optimization

The models used were SVM, Random Forest, Logistic Regression, XGBoost, Neural Network Classification and Convolutional Neural Networks.

Scalable Vector Machine (SVM): a supervised machine learning algorithm that helps you build the best dividing line (hyperplane) to separate the data points with the most space in between.

Random Forest: one of the ensemble learning methods in machine learning. It uses multiple decision trees to perform its predictions.

Logistic Regression: uses a sigmoid function to estimate how likely something belongs in a class based on its features

XGBoost (Extreme Gradient Boosting): one of the ensemble learning methods in machine learning. It uses sequential series of decision trees to perform its predictions. Each new tree corrects errors made by the previous ones.

Neural Network Classification: A neural network consists of layers of connected processing units (neurons) that work together to learn the best way to classify data.

Convolutional Neural Network: This is a neural network that automatically discovers the important features of data, through adjusting special filters.

E. Results and Discussion

Class	Precision	Recall	F1-Score	Support
0	0.97	0.96	0.96	67
1	0.94	0.96	0.95	47
Accuracy			0.96	114
Macro avg	0.95	0.96	0.95	114
Weighted avg	0.96	0.96	0.96	114

TABLE I: Classification using XGBoost

Class	Precision	Recall	F1-Score	Support
0	0.99	0.99	0.99	67
1	0.98	0.98	0.98	47
Accuracy			0.98	114
Macro avg	0.98	0.98	0.98	114
Weighted avg	0.98	0.98	0.98	114

TABLE II: Classification using SVM

Class	Precision	Recall	F1-Score	Support
0	0.98	0.96	0.97	67
1	0.94	0.98	0.96	47
Accuracy			0.96	114
Macro avg	0.96	0.97	0.96	114
Weighted avg	0.97	0.96	0.97	114

TABLE III: Classification using Random Forest

Class	Precision	Recall	F1-Score	Support
0	1.00	0.96	0.98	67
1	0.94	1.00	0.97	47
Accuracy			0.97	114
Macro avg	0.97	0.98	0.97	114
Weighted avg	0.98	0.97	0.97	114

TABLE IV: Classification using ANN

Class	Precision	Recall	F1-Score	Support
0	0.98	0.96	0.97	67
1	0.94	0.98	0.96	47
Accuracy			0.96	114
Macro avg	0.96	0.97	0.96	114
Weighted avg	0.97	0.96	0.97	114

TABLE V: Classification using Logistic Regression

Class	Precision	Recall	F1-Score	Support
0	0.98	0.97	0.98	67
1	0.96	0.98	0.97	47
Accuracy			0.97	114
Macro avg	0.97	0.97	0.97	114
Weighted avg	0.97	0.97	0.97	114

TABLE VI: Classification using CNN

Each set of results is comprised of multiple metrics to evaluate the model's performance i.e. Precision, Recall, F1-Score, Accuracy.

Note:

Support: Support refers to the number of instances of each class in the test set.

Macro Average: The macro average takes into account the performance of each class in the test set and computes the average metric score for each class, and then takes the average of these scores.

Weighted Average: The weighted average computes the average metric score for each class, but takes into account the relative size of each class in the test set.

The results show that the SVM model is performing best, with an accuracy of 0.98 (98%) . The weighted average of the F1-score is also consistently high, indicating that the model is making both accurate positive and negative predictions. The results also show that the model is performing well for all classes, with a high macro average of 0.98.

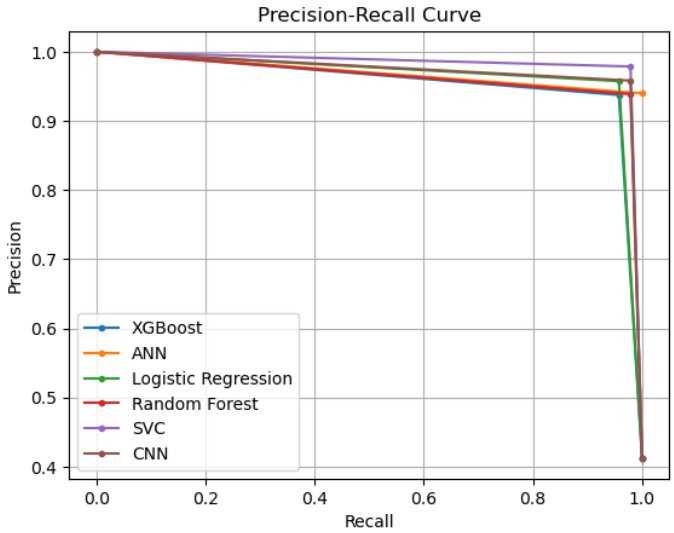

Fig. 3: Precision Recall curve

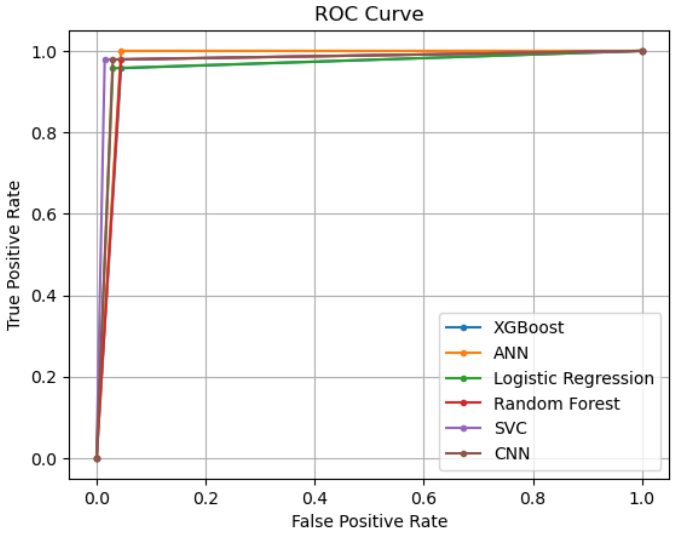

Fig. 4: ROC curve

SVM performs best with an area under the ROC curve of 0.982, followed by the ANN classifier with 0.978, then the CNN with 0.974.

F. AI model selection Accountability

AI accountability also involves ensuring that AI systems are transparent, understandable, and auditable, and that users have the ability to control and manage the data and information that is being used to train and operate AI systems. This can involve establishing clear policies and governance structures, creating guidelines for the development and deployment of AI systems, and ensuring that there is ongoing monitoring and evaluation to identify and address any unintended consequences or biases that may arise.

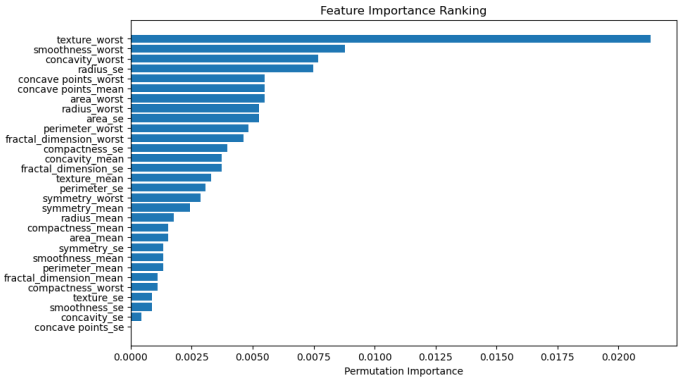

Fig. 5: Feature Importance

The concept of feature importance - the extent to which a certain feature influences the decision of a model - has been used to visualize the decision-making process of the model. As observed above, texture_worst greatly influences the positive classification of the model i.e. the presence of breast cancer.

III. Limitations

Grid search was performed on the models to find the best hyper-parameter for each model, but due to it's dynamic nature and reliance on computational capability, the inability to find static optimal hyper-parameters became evident.

The non-linear nature of SVM limits the number of explainable AI techniques that can be performed on the model to visualize the decision process.

IV. Conclusion and Future Works

In this concept paper, we proposed a breast cancer prediction model and experimented with six different models, with the Scalable Vector Machine classifier coming on top. Our experiments showed promising results, with an accuracy, macro average, and weighted average of 98% for precision, recall, and f1-score.

While our proposed system achieved promising results, there is still room for improvement. In future works, we plan to perform more explicit hyper-parameter tuning to further enhance the performance of the model. We also plan to display more explainable AI visualizations that aid in breaking down the black-box nature of the model.

In conclusion, we believe that our proposed breast cancer prediction model has the potential to aid health workers in their service towards fighting breast cancer. We look forward to further exploring and refining our system in future works.

V. Dataset and Code Links

Dataset Link:Wolberg,William, Mangasarian,Olvi, Street,Nick, and Street,W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.

Code Link:https://github.com/Emmanuel-Katwebaze/ML-project-coursework

References

[1] Develop, Bchd. (2023) Current breast cancer statistics and scenarios in Africa, Breast Cancer Hub. Available at: https://www.breastcancerhub.org/breast-cancer/about-breast-cancer-africa (Accessed: 25 April 2024).

[2] Shafi, A. (2023) Random Forest classification with Scikit-Learn, DataCamp. Available at: https://www.datacamp.com/tutorial/random-forests-classifier-python (Accessed: 25 April 2024).

[3] Eskandar, S. (2023) Introduction to RBF SVM: A powerful machine learning algorithm for non-linear data, Medium. Available at: https://medium.com/@eskandar.sahel/introduction-to-rbf-svm-a-powerful-machine-learning-algorithm-for-non-linear-data-1d1cfb55a1a (Accessed: 25 April 2024).

[4] GeeksforGeeks (2023) Gradient descent optimization in tensorflow, GeeksforGeeks. Available at: https://www.geeksforgeeks.org/gradient-descent-optimization-in-tensorflow/?ref=next_article (Accessed: 25 April 2024).

[5] XGBoost documentation (no date) XGBoost Documentation - xgboost 2.0.3 documentation. Available at: https://xgboost.readthedocs.io/en/stable/ (Accessed: 25 April 2024).

[6] Dash, S. (2023) Understanding the ROC and AUC intuitively, Medium. Available at: https://medium.com/@shaileydash/understanding-the-roc-and-auc-intuitively-31ca96445c02 (Accessed: 25 April 2024).