Utilizing Support Vector Machines for Breast Cancer Detection

Kirana Irfano
December 4th, 2024

Problem Statement

Breast cancer remains one of the most significant global health challenges, affecting millions of women and men each year. Despite recent advancements in medical science, breast cancer outcomes and survival rates may vary based on the early detection and accurate classification of the type of breast cancer. A crucial aspect of understanding breast cancer is distinguishing between its types, particularly malignant and benign tumors. This distinction is particularly important as it directly influences treatment decisions, prognosis, and the overall quality of life for patients. [2]

Malignant breast tumors are characterized by their aggressive nature and potential to invade surrounding tissues in certain parts of the body. This invasive behavior often requires intensive treatment regimens that include combinations of surgery, radiation, chemotherapy, and targeted therapies. The early identification of malignant tumors is crucial to mitigating the risk of widespread metastasis and improving patient outcomes. On the other hand, benign breast tumors, while non-cancerous, can still impact a patient's health and comfort. These tumors typically grow at a slower rate and do not invade surrounding tissues, making them less life-threatening. However, benign tumors may still cause some discomfort and may require removal if they grow too large and press against other organs in the body. Surgical removal may be required if benign tumors grow large enough to interfere with nearby organs or structures. [3]

The task of distinguishing between malignant and benign breast tumors is critical to ensuring timely and appropriate medical intervention. However, this process can be challenging due to overlapping clinical and imaging features between benign and malignant diagnoses. [3] Traditional diagnostic methods, while effective, may be time-consuming, resource-intensive, and prone to variability in interpretation. This emphasizes the need for reliable computational approaches to support medical decision-making and reduce diagnostic errors.

Solution

To address the challenges associated with the accurate classification of breast cancer tumors, this project focuses on leveraging machine learning techniques, specifically Support Vector Machines (SVM), to predict tumor types based on diagnostic data. By training the SVM model on a dataset containing key clinical and biological characteristics classifying a tumor related to breast cancer, the aim of this project is to create a tool that can accurately classify a tumor as malignant or benign. This computational approach increases the precision of the diagnosis, reducing the reliance on subjective interpretations and computationally inefficient conclusions. This model should ultimately provide healthcare professionals with actionable insights quickly and accurately.

The dataset used in this study contains a set of features derived from breast cancer diagnostics, including measurements such as the mean, standard error, and worst-case values of tumor radius, as well as the perimeter, area, compactness, concavity, and concave points of such tumors [1]. These features have been preprocessed and selected based on their correlation with the target variable (diagnosis: benign or malignant), ensuring that the most informative variables are used for classification. The analysis includes the following steps:

- Data Preprocessing: Removing redundant or empty values, and normalizing and standardizing each feature column.
- Feature Selection: Identifying and selecting diagnostic features that are most correlated with tumor classification.
- Model Training: Training an SVM classifier on the selected features and using cross-validation to ensure that the model does not under or overfit to the training data ultimately ensuring robustness.
- Model Evaluation: Evaluating the classifier's performance using metrics such as accuracy, precision, recall, and F1 score.
- Visualization: Generating classification reports, and confusion matrices to visualize results and provide transparency on the model's performance.

The ultimate goal of this project is to develop a high-performing SVM classifier capable of distinguishing between malignant and benign tumors with high accuracy. By integrating computational tools with traditional diagnostic methods, this approach aims to advance the field of computational oncology, improve patient outcomes, and contribute to the ongoing efforts to combat breast cancer.

Results and Demo

To view a live video demo, please see the **demo.mp4** video in the google drive folder linked in the resources section below.

Before training the model, the data was preprocessed by dropping irrelevant or redundant columns, and removing values that were null or empty. For each feature, outliers were removed by computing the IQR range for that specific column and removing any points that lie outside this range. Finally, each column was standardized and normalized (Figure 1).

```
# Standardization
scaler = StandardScaler()
df[column_names] = scaler.fit_transform(df[column_names])

# Normalization
normalizer = MinMaxScaler()
df[column_names] = normalizer.fit_transform(df[column_names])
```

Figure 1: Data Preprocessing

Once the data has been cleaned and preprocessed, feature selection was performed by computing the correlation matrix between all the features in the original dataset and the target feature (Figure 2). After

the correlation matrix was computed, features with a correlation coefficient above 0.5 with the target feature were chosen including: 'radius_mean', 'perimeter_mean', 'area_mean', 'compactness_mean', 'concavity_mean', 'concave points_mean', 'radius_se', 'perimeter_se', 'area_se', 'radius_worst', 'perimeter_worst', 'area_worst', 'compactness_worst', 'concavity_worst', 'concave points_worst'

	Correlation
concave points_worst	0.793566
perimeter_worst	0.782914
concave points_mean	0.776614
radius_worst	0.776454
perimeter_mean	0.742636
area_worst	0.733825
radius_mean	0.730029
area_mean	0.708984
concavity_mean	0.696360
concavity_worst	0.659610
compactness_mean	0.596534
compactness_worst	0.590998
radius_se	0.567134
perimeter_se	0.556141
area_se	0.548236

Figure 2: Correlation Matrix

Now that the data has been cleaned and the best features are selected, the model can be trained. To improve the robustness of the model and prevent both underfitting and overfitting, cross validation with 5 folds was used. For each fold, the cross-validation scores varied from 0.879 to 0.967 with an average cross validation score of 0.903. This means that the model trained with good accuracy using 5 different subsets of the training data. After fitting the model and predicting the training data the accuracy came out to be roughly 0.91, while the testing accuracy came out higher at 0.94. The classification report of the model's performance on the testing data is shown in Figure 3 below. Finally, the confusion matrix shown in Figure 4 demonstrates that the model achieves a high level of accuracy in distinguishing between benign and malignant tumors.

Classification Report:

	precision	recall	f1-score	support
0	0.92	1.00	0.96	71
1	1.00	0.86	0.93	43
accuracy			0.95	114
macro avg	0.96	0.93	0.94	114
weighted avg	0.95	0.95	0.95	114

Figure 3: Classification Report

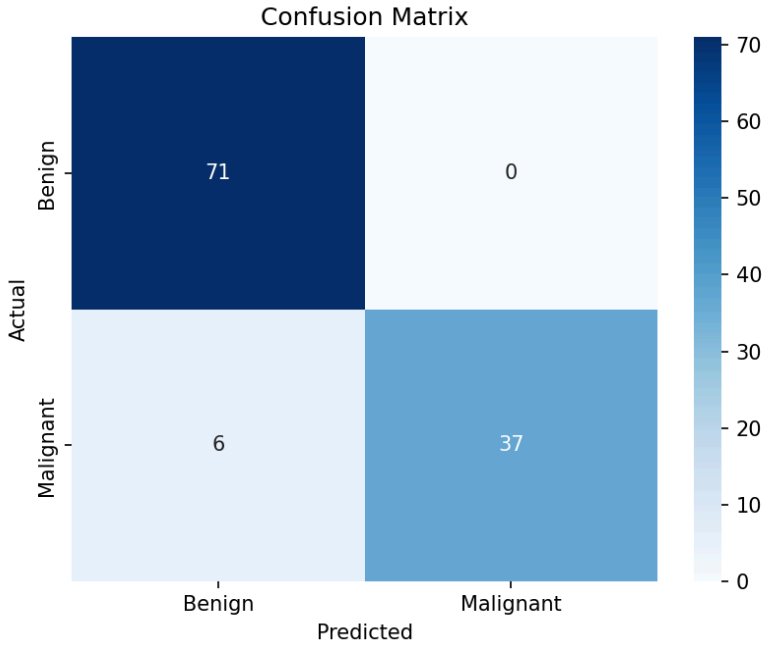

Figure 4: Confusion Matrix

These results indicate that, given sufficient and accurate measurements of the selected features used to train the model, healthcare providers can rely on the model to make quick and reliable predictions. Such efficiency and accuracy have the potential to aid in early detection and improve treatment planning, ultimately benefiting patient outcomes.

Assumptions, Constraints, and Implications

A large assumption in the process of training this model was made when selecting which features to use to train the model. In this implementation, a correlation matrix was used to determine which features were most relevant to the model based on its relationship with the target variable. It is assumed that the selected features (radius mean, perimeter mean, etc.) are sufficient and representative for accurately classifying the tumors as benign or malignant. However, there are other methods for feature selection which can produce slightly different and possible more accurate results which were not tested in this project.

Data availability is a large constraint in this project. The effectiveness of the model is constrained by the size and diversity of the dataset. This model may only be accurate for a certain demographic of people and works poorly to people with a different demographic from the people used for the dataset. Additionally, outlier removal using the IQR method assumes that all extreme values are irrelevant. This might nt always be the case and should a patient have a particularly large tumor, they should still be accurately diagnosed with a benign or malignant tumor. Finally, binary classification assumes a single decision threshold to separate malignant and benign tumors, however, there may be cases with uncertain or borderline classifications or diagnoses.

Some implications of the solution that should be made aware is the potential for the model to make errors. While this model has high levels of accuracy, it is not always correct. Healthcare providers should exercise caution when relying on the model's predictions as there is a risk of misclassification. For instance, the model might incorrectly classify a malignant tumor as benign. This incorrect diagnosis can put the patient at higher risk if they are given the wrong set of treatments.

Summary

Breast cancer poses a significant global health challenge, with accurate and timely classification of tumors being crucial for effective treatment. This project addresses this challenge by developing a SUpport Vector Machine model to classify breast tumors as either benign or malignant based on diagnostic data. Key features related to tumor characteristics including size, shape, and compactness were used to train the model.

The model demonstrates high accuracy, achieving a training accuracy of 91% and testing accuracy of 94%. The model's performance can be backed up by cross-validation, classification reports, and confusion matrices. These results suggest that, with sufficient , accurate, and diverse data, the model can assist healthcare providers by delivering rapid and reliable predictions. However, the model's effectiveness depends on several assumptions such as which features are selected to train the model and the data used.

While the solution can aid in early detection and treatment planning, caution is necessary when relying on predictions, as misclassifications could lead to significant risks for patients. Therefore, the SVM model should be used as a supplementary tool to assist, rather than replace the clinical judgement and procedures.

Links and Resources:

[1] https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset
[2] https://my.clevelandclinic.org/health/diseases/3986-breast-cancer
[3] https://www.cancercenter.com/community/blog/2023/01/whats-the-difference-benign-vs-malignant-tumors
[4] Link to Google Drive:
https://drive.google.com/drive/folders/1kWaajywMrBSUV2I0puqhVs-L9x_z2Uqa?usp=drive_link