CAMARINES SUR
POLYTECHNIC COLLEGES
COLLEGE of
COMPUTER STUDIES

Optimizing Breast Cancer Diagnosis in the Philippines Using SVM: A Cost-Effective Method

Divino Franco R. Aurellano
College of Computer Studies
Camarines Sur Polytechnic Colleges
diaurellano@my.cspc.edu.ph

Abstract

Breast cancer is a leading cause of death among women globally, particularly in the Philippines, where high healthcare costs and late diagnoses are common. As a computer science student, that serves as my motivation to develop an accessible and affordable method for early breast cancer detection using Support Vector Machine (SVM) algorithms. Utilizing the "Breast Cancer Wisconsin Dataset" from the UCI Machine Learning Repository, which contains 30 features from digitized breast tissue images. The GridSearchCV was employed method to optimize hyperparameters. It was determined that a linear kernel with C=0.0001 provided the best performance. The model achieved 97% accuracy, with a precision of 0.96, recall of 1.0, and F1-score of 0.98 for benign cases, and a precision of 1.0, recall of 0.93, and F1-score of 0.96 for malignant cases. These results demonstrate that this SVM model can effectively distinguish between malignant and benign tumors, potentially improving early breast cancer detection, reducing healthcare costs, and increasing survival rates, especially in resource-limited settings. Additionally, this model was deployed using Flask.

1 Introduction

Breast cancer is a disease characterized by the uncontrolled growth of abnormal breast cells, which form tumors. If untreated, these tumors can spread throughout the body and become fatal. This type of cancer remains as the leading cause of death among women worldwide, including in the Philippines. In 2020, it accounted for 685,000 deaths, or 15.5% of the total 4.4 million cancer deaths among women globally, and 9,926 deaths, or 21.8% of the 45,560 cancer deaths in the Philippines. The World Health Organization (WHO) report released on March 13, 2024, highlighted that there were 670,000 breast cancer-related deaths worldwide in 2022. Furthermore, a Philippine News Agency interview revealed that many women avoid getting checked-up due to financial concerns. According to WHO, the annual cost of cancer care in the Philippines is P35.3 billion.

As one of a student taking the field of computer science, my motivation to tackle this issue arises from the need to enhance early and accessible breast cancer diagnosis, and also the potential to reduce the

CSAC 224: Machine Learning 2024, Camarines Sur Polytechnic Colleges (LateX template borrowed from NIPS 2017.)

cost of patients as well for the healthcare provider. By this machine learning model, I developed an AI-based model using Support Vector Machine (SVM) algorithms to diagnose breast cancer. This model's input needed comes from computed digitized image of a fine needle aspirate (FNA) of a breast mass where the user will input the 30 various measurements like from perimeter_mean (mean size of the core tumor) and other required features. The model then predicts whether the tumor is malignant or benign. This approach is expected to improve early breast cancer detection, significantly increasing women's survival rates. Furthermore, I plan to implement this project in both rural and urban hospitals in the Philippines.

2 Related work

The related works in this section were found using a number of sources, including Google Scholar, National Institutes of Health, Science Direct, PLOS ONE,Semantic Scholar and others. Several studies have been conducted to investigate and implement machine learning algorithms in Diagnosing Breast Cancer. Most of them conducted a comparative study to know which learning algorithm is the most optimal to use in predicting breast cancer. Overall, the results show that SVM perform exceptionally well for breast cancer prediction, providing high accuracy and reliability.

Barrios (2022) analyzed the global challenges in breast cancer detection and treatment, discussing the high costs and poor outcomes due to late diagnosis. The study highlighted that screening mammography requires significant investment and is difficult to implement in resource-constrained settings. Barrios advocated for the adoption of new methods, such as machine learning algorithms, as a more effective and feasible solution to improve early detection and treatment outcomes.

Naji et al. (2021) explored various supervised machine learning approaches to determine the most effective algorithm for diagnosing breast cancer. They used SVM, Random Forest, Logistic Regression, Decision Tree, and K-Nearest Neighbors on the Breast Cancer Wisconsin Diagnostic Dataset. Their study concluded that SVM outperformed all other algorithms, achieving the highest accuracy of 97.2%, demonstrating the effectiveness of using SVM for breast cancer prediction.

Asri et al. [2016] in addition, this study, analyzes the most effective learning algorithm to be used in Breast Cancer Risk Prediction. By comparing the result of evaluation metrics by using accuracy, precision, sensitivity, and specificity. The results show that SVM gives the highest accuracy which is 97.13% and with lowest error rate. This also shows the importance of using comprehensive evaluation metrics to determine the best-performing algorithms in breast cancer risk prediction.

Islam et al.[2017] presented a novel modality to predict modality for the prediction of breast cancer, introducing Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) as supervised machine learning techniques. The Wisconsin Breast Cancer Diagnosis dataset from the UCI Machine Learning Repository was used in their study. Their approach provided better results for both training and testing phases. Specifically, the techniques achieved an accuracy of 98.57% and 97.14% for SVM and KNN, respectively, along with a specificity of 95.65% and 92.31% in the testing phase.

Aamir et al. [2022] developed a prediction model using a combination of Support Vector Machine (SVM) and Extremely Randomized Trees (Extra-Trees) classifier to diagnose breast cancer at an early stage based on risk factors. The Extra-Trees classifier was used for feature selection to remove irrelevant features, while SVM was utilized for diagnosis. They applied normalization to ensure that the data was well-distributed, which increased the success rate of the model. The model was evaluated using a breast cancer dataset with 116 subjects and stratified 10-fold cross-validation. Their proposed combined SVM and Extra-Trees model achieved an accuracy of 80.23%, significantly better than other machine learning models. This combined approach, with the normalization step, is expected to enhance diagnostic decision-support systems for breast cancer prediction.

Huang et al. (2017) conducted an analysis on the performance of Support Vector Machine (SVM) algorithms in predicting breast cancer, focusing on the impact of different kernel functions. Their experimental results indicated that for small-scale datasets, SVM ensembles with a linear kernel using the bagging method and SVM ensembles with a radial basis function (RBF) kernel using the boosting method were the most effective. They also emphasized the importance of feature selection during the data preprocessing stage for these datasets. In contrast, for large-scale datasets, SVM ensembles with an RBF kernel and boosting method outperformed other classifiers, demonstrating superior predictive performance.

You et al. (2010) conducted a comparative study of various classification techniques on breast cancer FNA biopsy data, specifically focusing on SVM, Bayesian classifiers, and other artificial neural network classifiers using the Wisconsin breast cancer dataset. They concluded that SVM outperformed the Bayesian network, offering higher prediction accuracy. Additionally, they evaluated the performance of these networks against other neural network approaches and found that the KNN algorithm achieved a 100% classification rate. Their findings underscored the effectiveness of these machine learning techniques, particularly SVM, in not only predicting breast cancer but also in diagnosing other challenging medical conditions.

Deshwat et al. (2019) presented a study in which they developed a model to predict breast cancer using Support Vector Machine (SVM) algorithms enhanced by GridSearchCV. Initially, they tested the SVM model without grid search and subsequently applied grid search to optimize the model. A comparative analysis was then conducted, leading to the construction of a new, more accurate model based on the grid search optimization. This study demonstrates the effectiveness of using grid search for parameter tuning, a technique also utilized in my research to improve model performance.

Khandelwal et al. (2023) In this recent study, authors investigated the application of various machine learning and AI algorithms for breast cancer detection using mammogram images. Their aim was to leverage open-source datasets and explore different breast cancer detection methodologies, including K-nearest neighbor (KNN), Convolutional Neural Networks (CNN), Support Vector Machines (SVM), and Generative Adversarial Networks (GAN), to identify the strengths and weaknesses of each approach. The study involved preprocessing and filtering the data and training multiple models on the Wisconsin dataset, starting with a basic logistic regression classifier. To gain a comprehensive understanding of the dataset and the parameters useful for training the models, they also performed data visualization. Ultimately, the authors found that the SVM model outperformed the KNN model, achieving a mean accuracy of 90% on the training data compared to 80% accuracy for KNN. This led to the conclusion that the SVM model offers superior performance and accuracy in this context.

Thombare et al. (2022) conducted a study to compare the efficiency and effectiveness of various machine learning classifier algorithms in predicting breast cancer using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. To facilitate practical application, they integrated their machine learning model into a web application using Flask and Firebase, allowing users to input values for 30 features and receive a breast cancer diagnosis. The Flask library played a crucial role in developing the project and ensuring seamless interaction between the model and the web interface.

3 Dataset and Features

The dataset I used in this study is the “Breast Cancer Wincosin Dataset” derived from the UCI Machine Learning Repository and was last updated 6 years ago. This contains 30 various features extracted from digitized images of breast tissue, which describes the characteristics of the cell nuclei from the image. Some features are radius_mean (mean of distances from center to points on the perimeter), texture ((standard deviation of gray-scale values) perimeter_mean (mean size of the core tumor) and 27 others. While there are 569 rows, 80% of it are for the training set and 20% are for the test set. Moreover, the features were normalized to ensure that they have a mean of 0 and a standard deviation of 1. This is crucial for SVM as it ensures that all features contribute equally to the decision boundary.

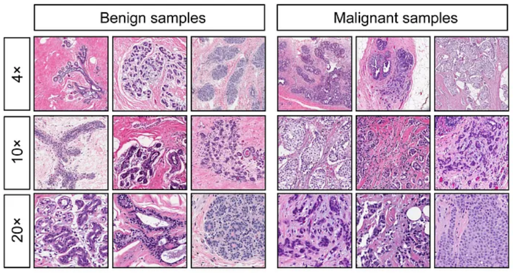


Figure 1: Sample image of a benign and malignant tumor cell

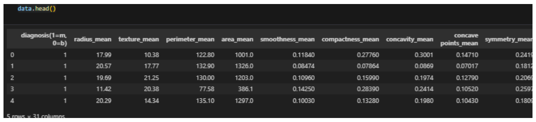
4 Methods

The learning algorithm I used in this model is Support Vector Machine, a very powerful tool and is widely used for classification and regression task. Where it often gives a cleaner and optimized way of learning complex for linear and non-linear data, through kernels. SVM works by finding the hyperplane that best separates data points of different classes in a high-dimensional space.The goal is to maximize the margin, which is the distance between the hyperplane and the nearest data points from each class, known as support vectors.

$$\min_{\theta} C \sum_{i=1}^{m} \left[y^{(i)} cost_1(\theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T x^{(i)}) \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2$$

Figure 2: Cost Function equation for SVM by Andrew Ng.

Before diving into model implementation, I conducted thorough Exploratory Data Analysis (EDA) which is part of my convention to ensure that the data I will be using is well-understood, clean, and appropriately prepared for modeling, leading to more reliable and accurate results. Following EDA, I recognized the necessity of data normalization. Subsequently, hyperparameter tuning was performed using GridSearchCV to identify the optimal combination of the regularization parameter (C) and kernel type. With the optimal hyperparameters determined, I proceeded to train the SVM model on the training data. And lastly, model evaluation was conducted, employing accuracy, precision, recall, and F1-score as primary metrics. Additionally, I utilized a confusion matrix to visually assess the model's performance by illustrating the distribution of correct and incorrect predictions.

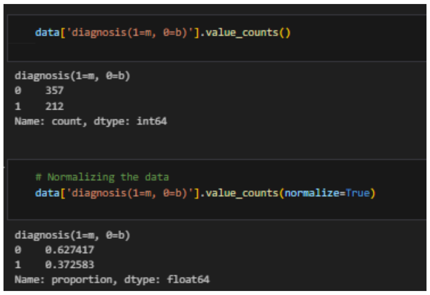


```
data.head()
```

	diagnosis(1=m, 0=b)	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
0	1	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419
1	1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812
2	1	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069
3	1	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597
4	1	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809

5 rows × 31 columns

Figure 3: After feature are computed in the image



```
data['diagnosis(1=m, 0=b)'].value_counts()
```

```
diagnosis(1=m, 0=b)
0    357
1    212
Name: count, dtype: int64
```

```
# Normalizing the data
data['diagnosis(1=m, 0=b)'].value_counts(normalize=True)
```

```
diagnosis(1=m, 0=b)
0    0.627417
1    0.372583
Name: proportion, dtype: float64
```

Figure 4: Sample part of EDA.

5 Experiments/Results/Discussion

Selecting the right hyperparameters is essential for the performance of the SVM model. I experimented with both linear and RBF (Radial Basis Function) kernels, ultimately choosing the linear kernel for its simplicity and interpretability. Various values for the regularization parameter C and the gamma parameter were tested. The GridSearchCV method was used to find the best combination of these hyperparameters. The best model was found to use that the best parameters were (kernel='linear', C=0.0001). The model was evaluated using primary metrics such as accuracy, precision, recall, and F1-score. The confusion matrix and classification report for the best SVM model on the test set indicated strong performance. For the benign diagnosis (case 0), the test classification report below showed a precision of 0.96, recall of 1.0, and an F1-score of 0.98, based on 71 samples. For the malignant diagnosis (case 1), the results were a precision of 1.0, recall of 0.93, and an F1-score of 0.96, based on 43 samples. The overall model accuracy was 97%. To check for overfitting, I compared the model's performance on the training set with the test set. The performance metrics for both sets were very close, indicating that the model generalizes well and is not overfitting. This also suggests that the model can make accurate predictions on new, unseen data.

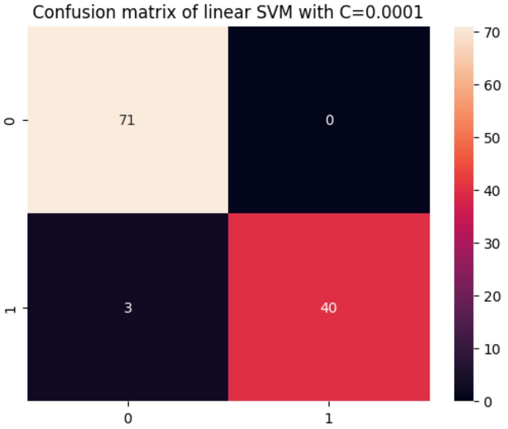


Figure 5: Confusion Matrix for test set

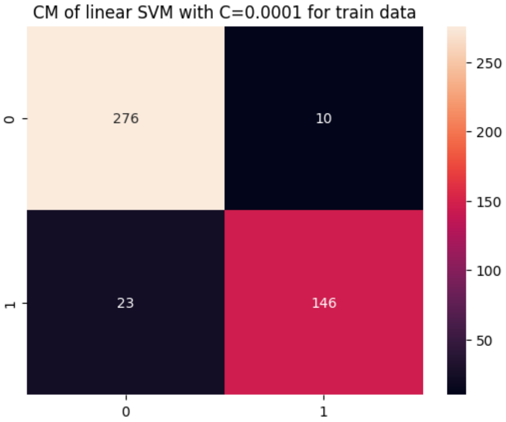
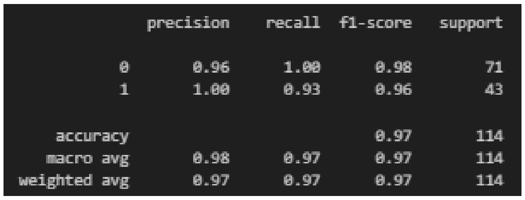
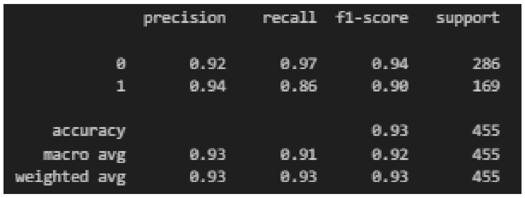


Figure 6: Confusion Matrix for train set



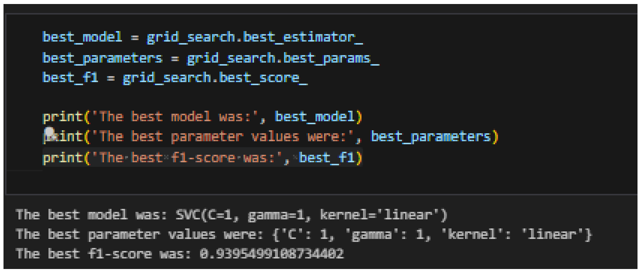
```
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        71
           1       1.00      0.93      0.96        43

    accuracy                           0.97       114
   macro avg       0.98      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114
```

Figure 7: Test set classification report



```
              precision    recall  f1-score   support

           0       0.92      0.97      0.94       286
           1       0.94      0.86      0.90       169

    accuracy                           0.93       455
   macro avg       0.93      0.91      0.92       455
weighted avg       0.93      0.93      0.93       455
```

Figure 8: Train set classification report



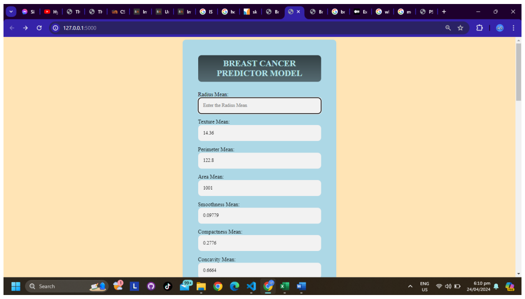
```
best_model = grid_search.best_estimator_
best_parameters = grid_search.best_params_
best_f1 = grid_search.best_score_

print('The best model was:', best_model)
print('The best parameter values were:', best_parameters)
print('The best f1-score was:', best_f1)
```

```
The best model was: SVC(C=1, gamma=1, kernel='linear')
The best parameter values were: {'C': 1, 'gamma': 1, 'kernel': 'linear'}
The best f1-score was: 0.9395499108734402
```

Figure 9: Gridsearchcv result

5.1 Model Deployment with Flask

In this section, the pickle library was utilized to save the trained model for deployment. The Flask framework was then employed to create a web application, enabling users to interact with the model. Users can input the values for the 30 features required by the model and receive a prediction on whether the breast cancer diagnosis is malignant or benign. Below are screenshots showcasing the user interface of the deployed model using Flask, demonstrating the functionality of the application.



127.0.0.1:5000

BREAST CANCER PREDICTOR MODEL

Radius Mean: Enter the Radius Mean

Texture Mean: 14.36

Perimeter Mean: 122.8

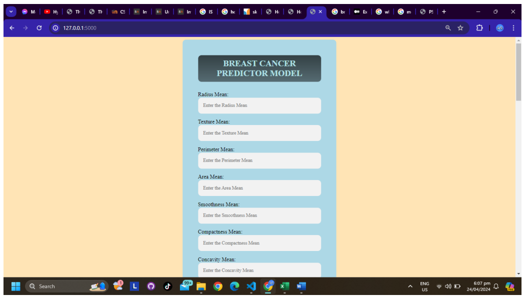
Area Mean: 1001

Smoothness Mean: 0.09779

Compactness Mean: 0.2776

Concavity Mean: 0.6664

Figure 10: Inputting features



127.0.0.1:5000

BREAST CANCER PREDICTOR MODEL

Radius Mean: Enter the Radius Mean

Texture Mean: Enter the Texture Mean

Perimeter Mean: Enter the Perimeter Mean

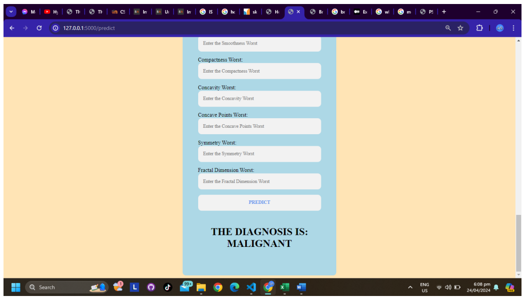
Area Mean: Enter the Area Mean

Smoothness Mean: Enter the Smoothness Mean

Compactness Mean: Enter the Compactness Mean

Concavity Mean: Enter the Concavity Mean

Figure 11: Inputting features



127.0.0.1:5000/predict

Enter the Smoothness Worst

Compactness Worst: Enter the Compactness Worst

Concavity Worst: Enter the Concavity Worst

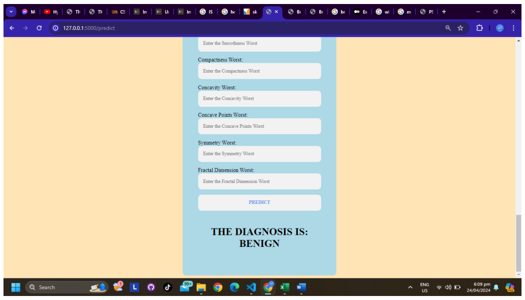
Concave Points Worst: Enter the Concave Points Worst

Symmetry Worst: Enter the Symmetry Worst

Fractal Dimension Worst: Enter the Fractal Dimension Worst

THE DIAGNOSIS IS: MALIGNANT

Figure 12: Prediction Result



127.0.0.1:5000/predict

Enter the Smoothness Worst

Compactness Worst: Enter the Compactness Worst

Concavity Worst: Enter the Concavity Worst

Concave Points Worst: Enter the Concave Points Worst

Symmetry Worst: Enter the Symmetry Worst

Fractal Dimension Worst: Enter the Fractal Dimension Worst

THE DIAGNOSIS IS: BENIGN

Figure 13: Prediction Result

6 Summary

This project aimed to develop an accessible and cost-effective method for early breast cancer detection using SVM, focusing on the Philippines where high healthcare costs and late diagnoses are common. Utilizing the "Breast Cancer Wisconsin Dataset" from the UCI Machine Learning Repository, which includes 30 features from digitized breast tissue images, a linear kernel SVM with C=0.0001 was found to perform best after hyperparameter optimization using GridSearchCV. The model achieved an accuracy of 97%, with a precision of 0.96, recall of 1.0, and F1-score of 0.98 for benign cases, and a precision of 1.0, recall of 0.93, and F1-score of 0.96 for malignant cases. Demonstrating its effectiveness in distinguishing between malignant and benign tumors. The linear SVM algorithm outperformed others due to its simplicity, robustness, and ability to handle high-dimensional data effectively. Furthermore, the model was successfully deployed using Flask. If I had more time, more members, or more computational resources, I would also like to explore expanding the dataset to include more diverse samples and incorporating other advanced techniques like deep learning to enhance model performance. I would also want to explore extensive hyperparameter tuning and resembling the methods to boost accuracy.

References

[1] Breast cancer. 13 Mar. 2024, www.who.int/news-room/fact-sheets/detail/breast-cancer.

[2] Cantal-Albasin, Grace. “Breast cancer: Not necessarily a death sentence, but a costly battle.” RAPPLER, 13 Nov. 2023, www.rappler.com/nation/mindanao/breast-cancer-not-death-sentence-costly-battle.

[3] “ Breast Cancer Wisconsin [Diagnostic] - EDA - Analytics Vidhya - Medium.” Medium, 8 Jan. 2022, medium.com/analytics-vidhya/breast-cancer-diagnostic-dataset-eda-fa0de80f15bd.

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

[5] Barrios, Carlos H. "Global challenges in breast cancer detection and treatment." The Breast 62 (2022): S3-S6.

[6] Naji, Mohammed Amine, et al. "Machine learning algorithms for breast cancer prediction and diagnosis." Procedia Computer Science 191 (2021): 487-492.

[7]Asri, Hiba, et al. "Using machine learning algorithms for breast cancer risk prediction and diagnosis." Procedia Computer Science 83 (2016): 1064-1069.

[8] Islam, Md Milon, et al. "Prediction of breast cancer using support vector machine and K-Nearest neighbors." 2017 IEEE region 10 humanitarian technology conference (R10-HTC). IEEE, 2017.

[9] Aamir, Sanam, et al. "Predicting breast cancer leveraging supervised machine learning techniques." Computational and Mathematical Methods in Medicine 2022 (2022).

[10] Huang, Min-Wei, et al. "SVM and SVM ensembles in breast cancer prediction." PloS one 12.1 (2017): e0161501.

[11] You, Haowen, and George Rumbe. "Comparative study of classification techniques on breast cancer FNA biopsy data." (2010).

[12] Deshwal, Vishal, and Mukta Sharma. "Breast cancer detection using SVM classifier with grid search technique." International Journal of Computer Applications 975.8887 (2019).

[13] Khandelwal, Atul Mishra, et al. “Breast Cancer Detection Using ML” IJARIIE-ISSN(O)-2395-4396 (2023)

[14] Thombare, Swapnil, et al. "DEVELOPMENT OF WEB APPLICATION FOR BREAST CANCER DETECTION USING MACHINE LEARNING CLASSIFIER." (2022)

[15] Andrew NG. Part, V. "Support vector machines."Standford Engineering Everywhere