OPTIMIZING BREAST CANCER DIAGNOSIS IN THE PHILIPPINES USING SVM: A COST-EFFECTIVE METHOD

Divino Franco R. Aurellano,
diaurellano@my.cspc.edu.ph

Jane B. Cagorong,
jacagorong@my.cspc.edu.ph

ABSTRACT

Breast cancer has been a leading cause of death among women globally, particularly in the Philippines, where high healthcare costs and late diagnoses have been common (World Health Organization [WHO], 2023). This served as the researchers' motivation to develop an accessible and cost-effective method for early breast cancer detection using Support Vector Machine (SVM) algorithms. The researchers utilized the Breast Cancer Wisconsin Dataset from the UCI Machine Learning Repository, which contained 30 features derived from digitized breast tissue images. GridSearchCV was employed to optimize the hyperparameters, and it was determined that a linear kernel with C = 0.0001 provided the best performance. The model achieved an accuracy of 97%, with a precision of 0.96, a recall of 1.0, and an F1-score of 0.98 for benign cases; and a precision of 1.0, a recall of 0.93, and an F1-score of 0.96 for malignant cases. These results demonstrated that the SVM model effectively distinguished between malignant and benign tumors, potentially improving early breast cancer detection, reducing healthcare costs, and increasing survival rates, especially in resource-limited settings. Additionally, the model was deployed using Flask.

KEYWORDS:

Breast cancer detection; Support Vector Machine; Machine learning; Healthcare in the Philippines; Flask deployment; Cost-effective diagnosis

1. RATIONALE

Breast cancer was the most common cancer among women worldwide and remained a leading cause of death. In the Philippines, it was the most commonly diagnosed cancer among women. According to the World Health Organization (2024), breast cancer has become the most common type of cancer globally. In 2022, the Philippines reported over 100,000 new cancer cases, and breast cancer was the most common among women (International Agency for Research on Cancer, 2022). Many Filipinos faced problems with early cancer diagnosis and treatment because of the high cost of healthcare and lack of advanced medical equipment. Public hospitals often had limited resources, which caused delays in diagnosis and treatment (Cantal-Albasin, 2023).

Because of these challenges, early and affordable detection methods were urgently needed. This gap in healthcare access motivated the researchers to explore alternative ways to help diagnose breast cancer at an early stage. The researchers chose to study the use of machine learning, particularly the Support Vector Machine (SVM) algorithm, which had been proven effective in classifying medical data with high accuracy. Studies showed that SVM worked well in identifying patterns in breast cancer datasets, which could help predict whether a tumor was benign or malignant (Al-Jumeily et al., 2016). The researchers decided to use the Breast Cancer Wisconsin Dataset from the UCI Machine Learning Repository (UCI, 2024) to test how accurate SVM could be in detecting breast cancer. This study aimed to provide a low-cost, reliable tool that could support doctors in making faster and more accurate diagnoses. By doing so, the researchers hoped to reduce delays in treatment and improve survival rates, especially in areas with limited healthcare resources.

2. SIGNIFICANCE OF THE STUDY

The proposed study will be beneficial to the following:

Filipino Healthcare Providers. This study will assist healthcare providers in the Philippines by offering a cost-effective method for early breast cancer detection. Using machine learning models like Support Vector Machine (SVM), the proposed approach will help improve diagnostic accuracy, especially in areas with limited medical resources.

Public Health Sector. The study contributes to the public health sector by providing a low-cost solution for breast cancer detection. By improving access to early screening, this research can reduce the overall healthcare burden, leading to better resource allocation and reducing treatment delays.

Researchers. Researchers in the fields of healthcare and machine learning will find this study valuable as it combines both disciplines. It adds to the body of knowledge on how machine learning can be used in healthcare, offering new avenues for future research and innovations in medical diagnostics.

Future Researchers. This study will serve as a foundation for future researchers interested in applying artificial intelligence and machine learning to medical fields. It encourages further exploration into how technology can improve healthcare systems, particularly in developing countries like the Philippines.

3. DESIGN/PROCESS/OUTPUT OF THE STUDY

This section presents a comprehensive overview of the study, including the dataset used, the methodology followed, and the implementation details of the system. Each part is described clearly and precisely to provide a complete understanding of the innovation.

3.1 Dataset and Features

The dataset used in this study is the "Breast Cancer Wisconsin (Diagnostic) Dataset", which was derived from the UCI Machine Learning Repository and last updated six years ago. This dataset contains 569 rows (instances) and 30 numerical features that were extracted from digitized images of breast tissue. These features describe the characteristics of the cell nuclei present in the images.

Some notable features include:

- radius_mean – the mean of distances from the center to points on the perimeter;
- texture_mean – the standard deviation of gray-scale values;
- perimeter_mean – the mean size of the core tumor;
- and 27 additional attributes describing other geometric and textural properties.

To ensure balanced model training and evaluation, the dataset was divided into an 80% training set and a 20% test set. Additionally, all features were normalized using standard scaling techniques, ensuring a mean of 0 and a standard deviation of 1. This normalization process is especially important for Support Vector Machine (SVM), as it guarantees that all features contribute equally to the decision boundary during model training.

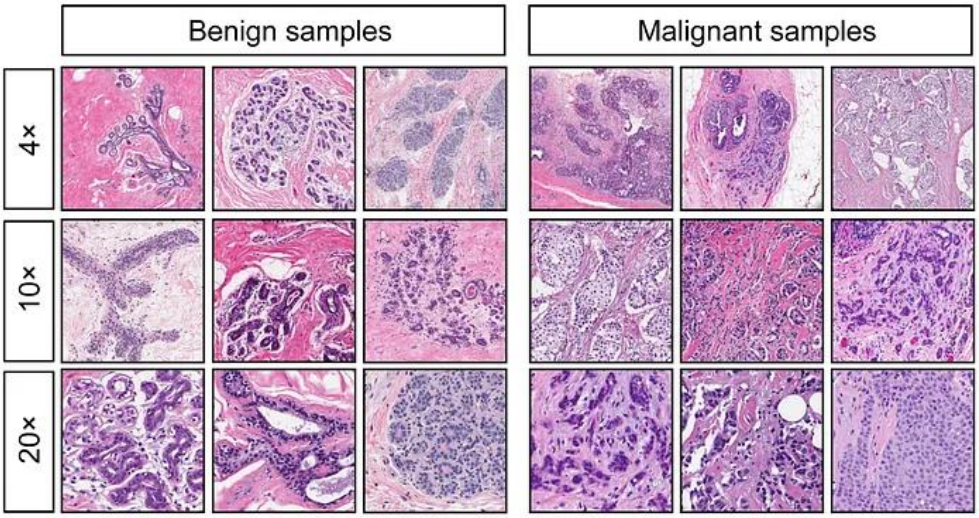


Figure 1: Sample image of a benign and malignant tumor cell

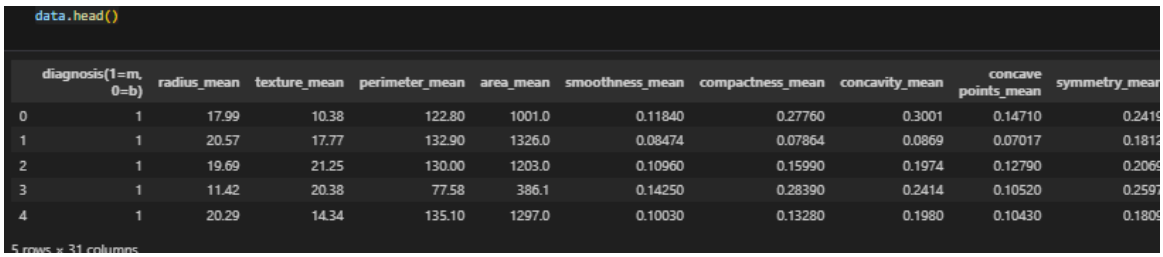
3.2 Methodology

The learning algorithm used in this study is the Support Vector Machine (SVM), a widely known supervised learning model used for classification and regression tasks. SVM works by identifying the optimal hyperplane that separates data points of different classes in a high-dimensional space. Its main objective is to maximize the margin—the distance between the hyperplane and the closest data points from each class, known as support vectors. Through kernel functions, SVM can handle both linear and non-linear classification tasks effectively.

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

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

Before implementing the model, we performed Exploratory Data Analysis (EDA) to understand and clean the dataset. This was followed by data normalization to ensure equal feature contribution. We then used GridSearchCV for hyperparameter tuning to find the best values for C and kernel type. After training the SVM model with the optimized parameters, I evaluated its performance using accuracy, precision, recall, F1-score, and a confusion matrix for a clear view 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

After EDA, feature normalization was performed to bring all features to the same scale. This ensures that each feature contributes equally to the decision-making process of the SVM model. Following this, hyperparameter tuning was conducted using GridSearchCV, a method that exhaustively searches through a specified parameter grid. The parameters tested include:

- Kernel types
- Regularization parameter (C)
- Gamma (for non-linear kernels)

```
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 4: GridSearchCV Result

Once the optimal parameters were selected, the model was trained using the training dataset (80% of the full data). After training, the model's performance was evaluated using the remaining 20% of the data (test set). Evaluation metrics included:

- Accuracy
- Precision
- Recall
- F1-score

- Confusion Matri

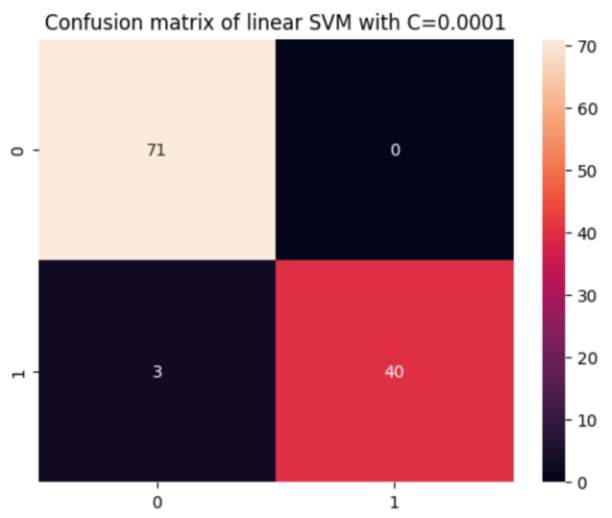


Figure 5: Confusion Matrix for test set

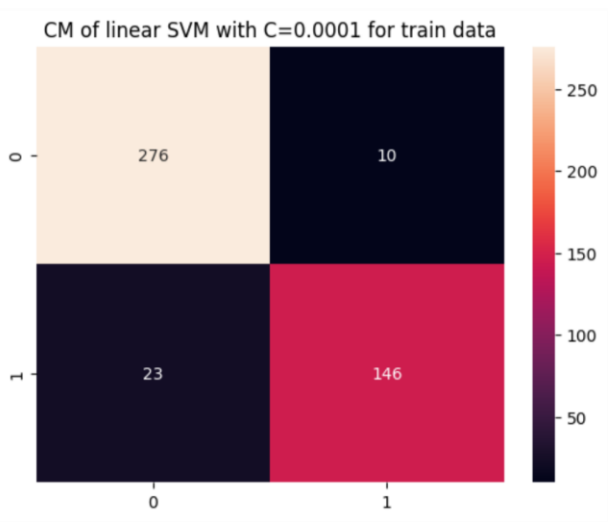


Figure 6: Confusion Matrix for train set

3.3 Implementation

To deploy the trained model, the Pickle library was used to serialize and save the final SVM model. For user interaction, the Flask web framework was employed to create a lightweight and accessible web application. This platform allows users to manually input the values of the 30 normalized features derived from breast tissue analysis.

Once the input form is submitted, the model processes the data and returns a prediction indicating whether the diagnosis is malignant or benign. This setup demonstrates how machine learning models can be integrated into real-world applications for clinical decision support.

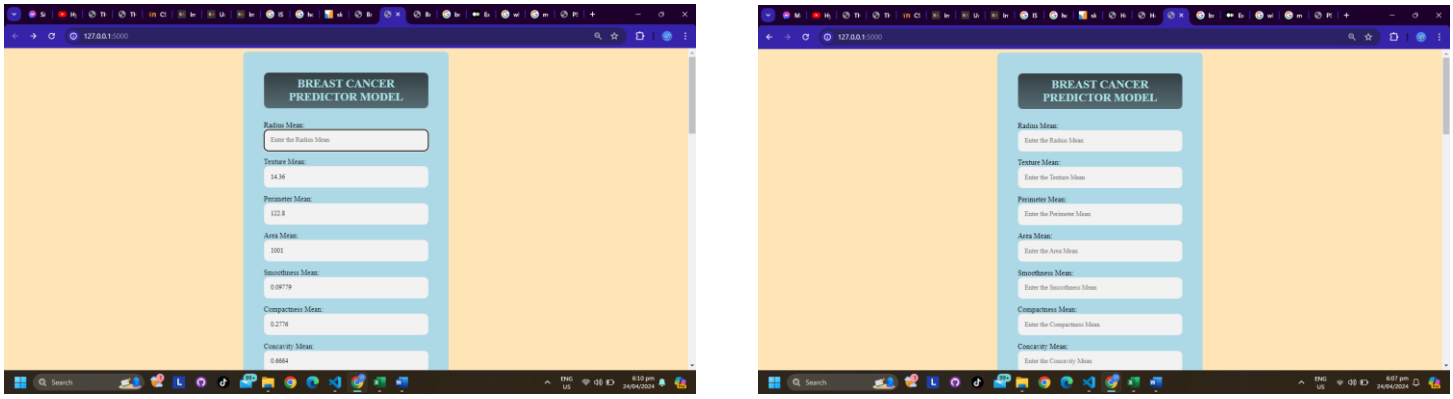


Figure 10: Inputting features

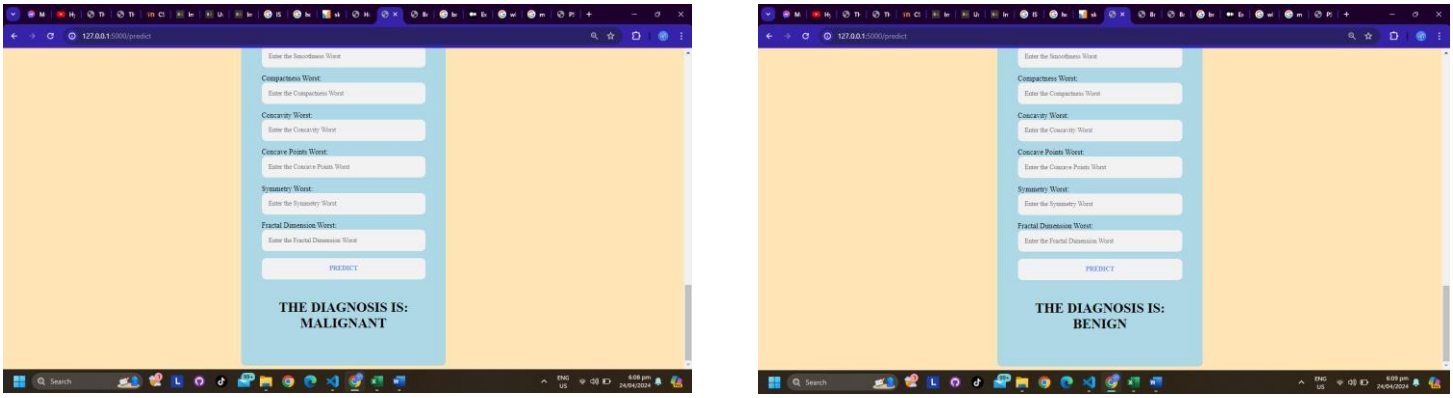


Figure 12: Prediction Result

4. FUTURE PLANS

The researchers plan to expand the dataset by incorporating more diverse samples to improve the model's accuracy and generalizability. Collaboration with healthcare providers, particularly in rural areas, will be pursued to gather real-world data. Future efforts will focus on exploring advanced machine learning techniques, such as deep learning, and optimizing the model through hyperparameter tuning and cross-validation. The Flask-based web application will also be enhanced for a better user experience and scalability, with potential integration into healthcare systems for real-time predictions.

5. REFERENCES

[1] Al-Jumeily, D., Hussain, A., Alghamdi, M., & Hamdan, H. (2023). *Breast cancer diagnosis using support vector machine optimized by meta-heuristic algorithms*. Scientific Reports, 13. https://www.nature.com/articles/s41598-024-61322-w

[2] International Agency for Research on Cancer. (2022). *Philippines fact sheet*. Global Cancer Observatory. https://gco.iarc.who.int/media/globocan/factsheets/populations/608-philippines-fact-sheet.pdf

[3] World Health Organization. (2024, March 13). *Breast cancer*. https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[4] Cantal-Albasin, G. (2023, November 13). *Breast cancer: Not necessarily a death sentence, but a costly battle*. RAPPLER. https://www.rappler.com/nation/mindanao/breast-cancer-not-death-sentence-costly-battle