Predicting Breast Cancer Within Cuban Women

Muhammad Razi Mahardika
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
muhammad.mahardika004@binus.ac.id

Marcell Kurniawan Sutanto
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
marcell.sutanto@binus.ac.id

***Abstract*—**This experiment focuses on evaluating a variety of classification model on predicting breast cancer within Cuban women between using all the features of the dataset and a select few features of the dataset decided by the importance of the features. The dataset includes 23 features that includes internationally recognized risk factors such as family history or breast cancer, lifestyle habits, demographic characteristics, and clinical outcomes. The machine-learning models used in this experiment are Random Forest Classification (RFC), Logistic Regression (LR), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). This study contributes to identifying reliable and accurate medical diagnosis tasks, specifically breast cancer detection.

***Keywords*—**Breast Cancer, Cuba, Binary Classification, Random Forest, Logistic Regression, KNN, SVM

I. Introduction

Breast cancer is the most common type of cancer that can be diagnosed within woman and the second leading cause of cancer death after lung cancer [1]. Breast cancer is defined as cells originating from breast tissues that proliferated or had an erratic growth [2]. People who are suffering from any types of cancer usually experience cancer related pain, depressive symptoms, and fatigue [3]. There are several probable risk factors for breast cancer such as aging, family history, reproductive factors, estrogen, and lifestyle [4].

Cases of breast cancer in Cuban women have shown to have a high incidence and high mortality rate with an average of 40 new cases per 100 thousand inhabitants [5]. The Cuban healthcare system has made a lot of developments and progress such as expanding the reach of its healthcare system to all its population, eradicating many types of diseases, improving its health indicators, and many other improvements [6]. Despite their accomplishments, Cuba still has some issues that needs to be addressed such as their increasing healthcare cost, inefficient allocation of resources, low hospital occupancy, and the deterioration of basic healthcare facilities [6]. It is important to consider the conditions of the healthcare system in a region as it may impact how cancer is diagnosed and treated. By diagnosing and treating breast cancer earlier, the risk of cancer death decreases significantly.

Many methods of diagnosing breast cancer have been developed over the years but most of them has its drawbacks. Mammography has been used as the standard breast cancer screening tool for its capability to significantly reduce breast cancer death, was found to be less effective when used to screen for women around the age of 42 and women with dense breast tissue [7,8]. MRI or magnetic resonance imaging is an alternative to mammography. MRI performs better at detecting breast lesions when compared to mammography but due to lacking specificity, may cause it to detect unwanted false positives [8]. Ultrasound has also been widely used as a complementary tool for breast cancer screening that covers the a few drawbacks of using mammography [9]. Several other methods of breast cancer diagnosis are Positron Emission Tomography (PET), microwave imaging, breast biopsy, and utilizing biomarkers [9].

With the advancement of technology, large amounts of data can be collected and analysed using machine learning. Models such as Neural Networks, Support Vector Machine (SVM), Decision Trees (DT), and K-Nearest Neighbour (KNN) are the usual choice of models used when diagnosing breast cancer using machine learning models [10]. With the help of machine learning models, doctors will be able to provide more accurate diagnosis on their patients, and with machine learning models being able to take over menial tasks for the doctor, the doctor will be able to save some effort and focus on much more important and complex tasks.

II. Literature Review

With such cancer being so common and dangerous to the female populace, it is important that a way to accurately classify and diagnose breast cancer is researched. This literature review section will focus on reviewing other works that has researched the same or similar cases close to breast cancer classification. Other research that didn't research breast cancer or similar cases but may have used methods that can be utilized for breast cancer classification will also be reviewed in this section.

For cases of breast cancer to get classified, machine learning models will need to be trained on a dataset before they are able to start classifying patients of breast cancer. For this, a study by J. M. Valencia-Moreno et al. [11] used a dataset containing a comprehensive dataset of Cuban women with detailed risk factors for breast cancer. This dataset is valuable for investigating region-specific predictors and allowed for the application of binary classification models. The study highlighted the importance of using epidemiological evidence to develop specialized diagnostic solutions. However, it did not show the performance of the different models on the dataset.

A study of breast cancer type classification by J. Wu et al. [12] worked on the classification of breast cancer type triple negative breast cancer (TNBC) and non-TNBC. They used many models, and four models were evaluated for their

performance, including, Support Vector Machine (SVM), K-nearest neighbour (KNN), Naive Bayes (NGB) and Decision Tree (DT). The study reveals that the model SVM can accurately classify breast cancer that are TNBC and non-TNBC. However, further research is recommended to identify the best features to be used.

Another research did a similar study for the classification of breast cancer using machine learning by N. Rane et al. [13]. The study aims to show a comparison of machine learning algorithms on classifying the cancer as benign or malignant. The models used were Naive Bayes (NB), Random Forest (RT), Artificial Neural Networks (ANN), Nearest Neighbour (KNN), Support Vector Machine (SVM) and Decision Tree (DT) on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset which is extracted from a digitised image of an MRI. However, the paper did not show any result of the performance of any models.

One research done by A. Bhardwaj et al. [14] tested several machine learning models on the same WDBC dataset with the objective of classifying benign and malignant tumours. The models they chose to test were Multilayer Perceptron (MLP), K-Nearest Neighbour (KNN), Genetic Programming (GP), and Random Forest (RF). The paper made a comparison on each of the model and found that the RF algorithm obtained the highest f1-score out of the four models with MLP having the lowest f1-score.

Research by M. Nawaz et al. [15] did a multi class breast cancer classification by using deep learning Convolutional Neural Network (CNN). For the dataset, they used the BreakHis [16] dataset for training and testing. It contains 7909 microscopic biopsy images of benign and malignant breast tumors. The model used for training and testing is the DenseNet model and they modified it to extract fully global feature. It achieved an accuracy of 96% which is slightly higher compared to a popular model mentioned in the paper, CSDCNN, which achieved an accuracy of about 94%. The result of their model of multi-class breast cancer classification task outperforms human expert in the diagnostic domain.

Another research by B. S. Abunasser et al. [16] used the Xception alogorithm for breast cancer detection and classification. The objective of the study is to propose a deep learning model for detecting and classifying breast cancer. The dataset used are BreakHis [16] dataset for training and testing. It contains 7909 microscopic biopsy images of benign and malignant breast tumors. The dataset was boosted with Generative Adversarial Network (GAN) to generate more images since they mentioned that the dataset size was considered low. The model achieved a precision score of 97.60%, recall score of 97.60% and F1 score of 97.58%.

III. Methodology

The methodology section can be split into the following sections described by the flowchart below.

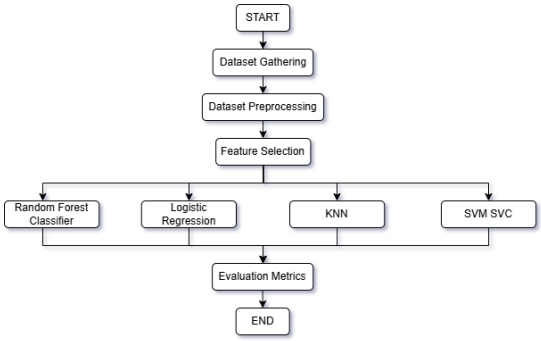

Fig. 1. Methodology Flowchart

The methodology section will start by obtaining the dataset. Once the dataset has been obtained, the data will be analysed for any issues that may affect the machine learning models. If any issues are identified, it will be resolved at the data preprocessing section. Once the dataset has been pre-processed, feature selection will be done on the dataset using several feature selection methods to identify the most significant feature for predicting the target label. Once the features are selected, each model chosen will be experimented on using the dataset that only has the chosen features and the dataset that still has all its features. This means that there will be a total of eight experiments being done. Once all eight experiments have been conducted, each of the experiment's performance will be evaluated using several evaluation metrics.

A. Dataset Gathering

The dataset that this research uses is the "Cuban Breast Cancer" dataset. The dataset contains data of patients from Cuba that consulted with a hospital in Havana for breast cancer. The dataset contains 23 variables used to represent the risk factors of breast cancer and 1697 data instances. The target label that will be used in this dataset is the "cancer" variable that includes boolean values "Yes" or "No".

B. Dataset Preprocessing

Preprocessing the dataset is a crucial step in machine learning that can help increase the machine learning model's performance and avoid having issues when training the machine learning model.

The dataset contains several issues that can hinder the machine learning model process. First, a lot of data instances still contain missing values and will need to be handled with. Second, some columns contain unique values that contains unnecessary suffixes such as "1 month" or "2 months" where the suffix "month(s)" can be removed so that the column can be turned into an integer datatype. Third, some columns contain multiple unique values that all represents the same meaning such as the string value "No" and the integer value 0 in some columns. Lastly, some data instances combine multiple unique values of a column to a single data instance. How the dataset combines multiple values is by appending additional values at the end of the previous string value separated by a slash (/). One example of this is in one of the columns, a string value "Mother/Sister" combines the unique value "Mother" and "Sister" into one new unique value and assigns it to a data instance. Because every new unique combination of values created in this manner can cause the

machine learning model to consider more unique values then they are supposed to, a new way to represent data instances with multiple values needs to be considered.

With the issues within the dataset identified, specific preprocessing methods can be utilized to resolve the issues. The preprocessing steps that were taken are as follows:

1. Replacing unique values that represents the same meaning with a uniformed value using replace functions.
2. Removing unnecessary suffixes using string replace function accompanied by regex syntax.
3. Creating new binary columns for columns that can contain a combination of values and assigning a "True" or "False" to the binary column by checking the string value for each unique value.
4. Handling missing values by:
    a. Removing the data instances with missing values if there aren't much missing values to handle in the column.
    b. Replacing missing values with the median of the column if the value of the data isn't significant.
    c. Imputing missing values with the results of the random forest regressor if the value of the data is significant.
5. Encoding categorical data into integers using one hot encoding or label encoding

C. Feature Selection

Each model being used in this experiment will be tested on two types of datasets. The first dataset will have all the features available in the original dataset, and the second will have only a few chosen features that were considered after seeing the results of three feature selection methods. The three feature selection methods that will be utilized are as follows:

1. Tree-based feature importance using random forest classifier.
2. Recursive feature elimination using logistic regression.
3. Select K-best using the chi2 score.

The results of each of the feature selection methods vary from each other. For this research, the most consistent top performing features across all methods will be chosen to be used for the second dataset. The selected features are: "biopsies", "histologicalclass", "consumed_alcohol", "menopause", and "is_sad".

D. Experimental Setup

Binary classification is one of the many tasks that can be achieved in machine learning. The goal of binary classification is to classify each data instance to one of two possible classes. This research is trying to classify its data instances to identify if a patient is suffering from breast cancer or not.

Several machine learning models can be used to perform this task. Each machine learning model has its own algorithms, advantages, and limitations. This research will be using four models and making a comparison between them to find which model is best suited for this task and dataset.

The first model is the random forest classifier. The random forest classifier works by creating many different decision trees which are trained on different parts of the dataset and conducts voting between all the trees to find the most likely predictions. This model is known to provide accurate predictions with large datasets and helps reduce the risks of overfitting. However, this model is limited due to the high computational cost to run it.

The second model is the logistic regression model. The logistic regression model is usually used for binary classification tasks. This model uses the sigmoid function that uses variables as inputs to produce a float value between 0 and 1. If the resulting value is closer to 0 then the data instance will be assigned to class 0, and such is the case for class 1. The logistic regression model is known to be extremely fast when classifying new data instances. The model is limited in its ability to capture complex relations within the dataset.

The third model is the K-nearest neighbour (KNN) model. The KNN model is a supervised learning model that can be used for both classification and regression tasks. This model works by considering the k amount of the nearest datapoints to the new datapoint that is trying to be classified and assigns a class to the new datapoint depending on what the majority population is among its neighbouring datapoints. KNN is widely used in machine learning due to its versatility, simplicity, and ease of use. However, KNN is prone to overfitting, does not scale well, and does not perform well when the dataset used has a high dimensionality.

The fourth model is the support vector machine (SVM) model, specifically the support vector classification (SVC) version of the model. SVM is widely used for linear and nonlinear classification tasks as well as regression tasks. SVM works by finding a maximum separating hyperplane between each class available in the target label. The separating hyperplane is created by trying to find the maximizing distance between the nearest datapoint of each class and the hyperplane itself. SVM models are known to perform well in high dimensional datasets and are memory efficient when performing their task due to their focus on support vectors. The model is limited due to its slow training time, sensitivity to noise, and feature scaling sensitivity.

All the above models will be trained, tested, and evaluated on the two types of datasets, the normal dataset and the dataset with only the most significant features. With four models, there is a total of eight experiments that will be conducted. The experiments are as follows:

TABLE I. EXPERIMENT LIST

Experiment No.	Model	Dataset
Experiment 1	Random Forest	Normal
Experiment 2	Random Forest	Best Features
Experiment 3	Logistic Regression	Normal

Experiment 4	Logistic Regression	Best Features
Experiment 5	KNN	Normal
Experiment 6	KNN	Best Features
Experiment 7	SVM SVC	Normal
Experiment 8	SVM SVC	Best Features

E. Evaluation Metrics

To identify if the models performed well, an evaluation needs to be performed using evaluation metrics. This research will be using four scoring metrics to help evaluate each experiment. These metrics include accuracy score, precision score, recall score, and f1-score.

Accuracy score is used to measure the general performance of the model by seeing how the model's outcome matches the actual outcome. The score is calculated by counting the number of correct predictions over the total predictions made.

Precision score measures how accurate are the positive predictions. It calculates the score by counting the amount of true positive predictions made over the total of true positives and false positives.

Recall score measures the model's ability to identify true positive cases. It calculates the score by counting the amount of true positives predictions over the total of true positives and false negatives.

F1-score is used to measure the balance between precision and recall scores. For a model to receive a high f1-score, the model would need to achieve a high performance on both precision and recall score. Equation (1) shows how the f1-score is calculated.

$$2 * \frac{Precision * Recall}{Precision + Recall} \quad (1)$$

IV. Results and Discussion

This section will show the performance of each of the model in each experimental setups and make a comparison based on the metrics that was decided. The results of experiment 1 and 2 which uses Random Forest as its model can be seen on table II and table III.

TABLE II. Experiment 1 results

Experiment No.	Precision Score	Recall Score	F1-Score
Positive	1.00	1.00	1.00
Negative	1.00	1.00	1.00
Macro Avg	1.00	1.00	1.00
Weighted Avg	1.00	1.00	1.00
Accuracy			1.00

TABLE III. Experiment 2 results

Experiment No.	Precision Score	Recall Score	F1-Score
Positive	1.00	0.99	0.99
Negative	0.97	0.99	0.98
Macro Avg	0.99	0.99	0.99
Weighted Avg	0.99	0.99	0.99
Accuracy			0.99

From the results, the random forest model was able to achieve a perfect performance on all metrics when fitted using all available features. With perfect scores across all metrics, it is possible that the model may have overfitted during training. When trained on the dataset that only has the selected features, it still has a high performance but has some imperfect results as seen by the metrics. It is possible that feature selection might have helped slightly with the overfitting problem.

Experiment 3 and 4 uses logistic regression for its experiment. The results can be seen on table IV and V below.

TABLE IV. Experiment 3 results

Experiment No.	Precision Score	Recall Score	F1-Score
Positive	1.00	1.00	1.00
Negative	1.00	1.00	1.00
Macro Avg	1.00	1.00	1.00
Weighted Avg	1.00	1.00	1.00
Accuracy			1.00

TABLE V. Experiment 4 results

Experiment No.	Precision Score	Recall Score	F1-Score
Positive	0.99	1.00	0.99
Negative	0.99	0.98	0.99
Macro Avg	0.99	0.99	0.99
Weighted Avg	0.99	0.99	0.99
Accuracy			0.99

The results of the experiments shows that the logistic regression model suffers the same issues as the random forest model in the previous experimental setup. When used with the complete dataset, it was able to achieve perfect scoring on all metrics and when used with the dataset that only has the selected features, it was able to nearly achieve perfect scoring. This could be a sign that overfitting also happened in the training of the logistic regression model.

For experiment 5 and 6, the K-Nearest Neighbour model was used. Its performance can be viewed in table VI and VII below.

TABLE VI. EXPERIMENT 5 RESULTS

Experiment No.	Precision Score	Recall Score	F1-Score
Positive	1.00	1.00	1.00
Negative	1.00	1.00	1.00
Macro Avg	1.00	1.00	1.00
Weighted Avg	1.00	1.00	1.00
Accuracy			1.00

TABLE VII. EXPERIMENT 6 RESULTS

Experiment No.	Precision Score	Recall Score	F1-Score
Positive	0.98	0.98	0.98
Negative	0.96	0.96	0.96
Macro Avg	0.97	0.97	0.97
Weighted Avg	0.98	0.98	0.98
Accuracy			0.98

The results show that the KNN model when trained on the complete dataset still has perfect performance like previous cases. When the KNN model was trained on the dataset with selected features, it has a slightly lower score when compared to previous experimental setups that used the dataset with selected features. However, the difference isn't significant enough to dismiss the possibility of overfitting.

Experiment 7 and 8 uses the Support Vector Classification version of the Support Vector Machine model. Its results can be seen in table VIII and IX below.

TABLE VIII. EXPERIMENT 7 RESULTS

Experiment No.	Precision Score	Recall Score	F1-Score
Positive	0.99	1.00	1.00
Negative	1.00	0.98	0.99
Macro Avg	1.00	0.99	0.99
Weighted Avg	0.99	0.99	0.99
Accuracy			0.99

TABLE IX. EXPERIMENT 8 RESULTS

Experiment No.	Precision Score	Recall Score	F1-Score
Positive	0.83	0.94	0.88
Negative	0.84	0.59	0.69
Macro Avg	0.83	0.77	0.79
Weighted Avg	0.83	0.83	0.82
Accuracy			0.83

The SVM SVC model has the lowest scores out of all the other models and the only model that didn't achieve perfect scores when trained on the complete dataset. When the model is trained on the feature selected dataset, it's able to achieve an accuracy score of 83%. Although the score is lower than the other models, it shows that the model is still capable of making mistakes, which could be an indication that the SVM SVC model was able to resolve the overfitting problem. This is also backed up by the fact that SVM models can work well with high dimensional data which could have contributed to it being able to prevent overfitting during training.

V. CONCLUSION

This research had investigated the performance of several machine learning models when used to classify possible cases of breast cancer. The models Random Forest, Logistic Regression, Support Vector Classification, and K-Nearest Neighbour were all experimented and trained on the Cuban Breast Cancer dataset. Using accuracy, precision, recall, and f1-score, it can be determined that many of the models may have had overfitted during training due to their perfect metric scores. With the exception being the SVM model when used with the feature selected dataset.

We acknowledge that there may be limitations within our studies. The dataset that was used was not normalized and still has many dimensions within it. In the future, if the same dataset was to be used again, it would be optimal to perform dimensionality reduction and data normalization before being used to train the machine learning models. With these additional steps, a better performing model may be created.

REFERENCES

[1] A. N. Giaquinto et al., 'Breast cancer statistics, 2022', CA: a cancer journal for clinicians, vol. 72, no. 6, pp. 524–541, 2022.

[2] G. N. Sharma, R. Dave, J. Sanadya, P. Sharma, and K. K. Sharma, 'Various types and management of breast cancer: an overview', Journal of advanced pharmaceutical technology & research, vol. 1, no. 2, pp. 109–126, 2010.

[3] D. Carr et al., 'Management of cancer symptoms: pain, depression, and fatigue', in Database of Abstracts of Reviews of Effects (DARE): Quality-assessed Reviews [Internet], Centre for Reviews and Dissemination (UK), 2002.

[4] Y.-S. Sun et al., 'Risk factors and preventions of breast cancer', International journal of biological sciences, vol. 13, no. 11, p. 1387, 2017.

[5] Y. G. Alvarez, L. F. Garrote, P. T. Babie, M. G. Yí, and M. G. Jordán, 'Breast cancer risk in Cuba', MEDICC Review, vol. 5, pp. 2–3, 2003.

[6] F. E. Sixto, 'An evaluation of four decades of Cuban healthcare', Cuba in Transition, vol. 12, pp. 325–343, 2002.

[7] B. N. Hellquist et al., 'Effectiveness of population-based service screening with mammography for women ages 40 to 49 years: evaluation of the Swedish Mammography Screening in Young Women (SCRY) cohort', Cancer, vol. 117, no. 4, pp. 714–722, 2011.

[8] D. Roganovic, D. Djilas, S. Vujnovic, D. Pavic, and D. Stojanov, 'Breast MRI, digital mammography and breast tomosynthesis: comparison of three methods for early detection of breast cancer',

Bosnian journal of basic medical sciences, vol. 15, no. 4, p. 64, 2015.

[9] L. Wang, 'Early diagnosis of breast cancer', Sensors, vol. 17, no. 7, p. 1572, 2017.

[10] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, 'Machine learning with applications in breast cancer diagnosis and prognosis', Designs, vol. 2, no. 2, p. 13, 2018.

[11] J. M. Valencia-Moreno, J. A. Gonzalez-Fraga, E. Gutierrez-Lopez, and H. A. Cantero-Ronquillo, 'A dataset of breast cancer risk factors in Cuban women: Epidemiological evidence from Havana', Data in Brief, vol. 57, p. 111029, 2024.

[12] J. Wu, and C. Hicks, 'Breast Cancer Type Classification Using Machine Learning', J. Pers. Med., vol. 11, Issue 2, 10.3990, jpm11020061, 2021.

[13] N. Rane, J. Sunny, R. Kanade, and Prof. S. Devi, 'Breast Cancer Classification and Prediction using Machine Learning', International Journal of Engineering and Research & Technology (IJERT), vol. 9, Issue 2, 2020.

[14] A. Bhardwaj, H. Bhardwaj, A. Sakalle, Z. Uddin, M. Sakalle, and W. Ibrahim, 'Tree-Based and Machine Learning Algorithm Analysis for Breast Cancer Classification', Computational Intelligence and Neuroscience, vol. 2022, no. 1, p. 6715406, 2022.

[15] M. Nawaz, A. A. Sewissy, T. H. A. Soliman, 'Multi-Class Breast Cancer Classification using Deep LLearning Convolutional Neural Network', International Journal of Advanced Computer Science and Application (IJACSA), vol. 9, no. 6, 2018.

[16] F.A. Spanhol, A. Fabio., L.S Oliveura., C. Ptitjean. A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering, 2016, vol. 63, no 7, p. 1455-1462.