Journal of Applied Engineering and Technological Science
Vol 5(1) 2023: 401-413

STANDARDSCALER'S POTENTIAL IN ENHANCING BREAST CANCER ACCURACY USING MACHINE LEARNING

Febri Aldi[1*], Febri Hadi[2], Nadya Alinda Rahmi[3], Sarjon Defit[4]
Information Technology Doctoral Program, Putra Indonesia University YPTK, Indonesia[1234]
febri_aldi@upiyptk.ac.id

Received : 17 August 2023, Revised: 16 November 2023, Accepted : 21 November 2023
**Corresponding Author*

ABSTRACT
The major consequence of breast cancer is death. It has been proven in many studies that machine learning techniques are more efficient in diagnosing breast cancer. These algorithms have also been used to estimate a person's likelihood of surviving breast cancer. In this study, we employed machine learning algorithms to predict breast cancer. The aim of this research is to increase accuracy in predicting breast cancer. A total of 569 breast cancer datasets were obtained from kaggle sites. Some of the machine learning algorithms that we use are K-Nearest Neighbor (KNN), Random Forest (RF), Gradient Boosting (GB), Gaussian Naive Bayes (GNB), Vector Support Machine (SVM), and Logistic Regression (LR). Before algorithms were used to train and test breast cancer datasets, StandardScaler was leveraged to transform training datasets and test datasets for improved algorithm performance. As a result of this utilization, the performance measurement carried out succeeded in producing high accuracy. The highest results were obtained from the Logistic Regression algorithm with an accuracy value of 99%. The value of precision is 99% benign, and 100% malignant. The recall results are 100% benign, and 98% malignant. The F1-Score results show 99% benign, and 99% malignant. It is hoped that this research can help the medical party to determine the next step in dealing with breast cancer.
Keywords: *Breast Cancer, Logistic Regression, Machine Learning, StandardScaler*.

1. Introduction

Breast cancer has been the cause of numerous fatalities. Every year, according to WHO, there are approximately 1.5 million cases of breast cancer that attack women worldwide. Breast carcinoma, one of the most well-known malignancies, was first discovered in Egypt around 1600 BC (Monirujjaman Khan et al., 2022). In Indonesia, the type of disease that often attacks the breasts of women resulting in death is breast cancer (Widiana & Irawan, 2020). 16.6% of Indonesia's 396,914 new instances of cancer, or 68,858 new cases, were breast cancer (Fadilah et al., 2022). More than 22,000 deaths were reported during this time. In fact, when patients regularly practice early detection and minimize cancer-causing risk factors, nearly 43% of cancer-related deaths are preventable. Breast cancer can be found through tumors. Malignant or benign tumors are classified as tumor types (Mekha & Teeyasuksaet, 2019). The doctor must use active determination strategies to find aggressive malignancies. But even for experts, it's quite difficult to detect cancer (Hughes et al., 2022). Therefore, automated cancer detection techniques are required. Often studies have tried to use Machine Learning (ML) techniques to predict a person's propensity to survive cancer. These algorithms appear to be more effective at detecting carcinoma (Nozomi et al., 2022). Usually, the accuracy of patient detection requires the experience and knowledge of the doctor (Chakraborty et al., 2019). However, these skills have been developed over the years to confirm diagnosis and observe the negative effects of many individuals. Even so, dependence cannot be guaranteed. Because processing technology has advanced (Klein et al., 2021).

Large amounts of data can now be collected and stored with relative ease, such as in specialized databases of patient data electronically (Cios & William Moore, 2002). Health professionals would not be able to decipher this huge database without the help of computers, especially when performing significant data analysis (van der Niet & Bleakley, 2021). Correctly classifying severe tumors can also keep some patients from receiving the required care (Tazin et al., 2021). Therefore, a contentious scientific issue is the precise diagnosis and classification of breast cancer into benign and malignant categories. ML approaches were widely utilized to recognize breast cancer and infers new ideas from data patterns in the last century. The use of

machine learning to categorize and model breast cancer is widely known (Amrane et al., 2018). Hidden patterns and regularities in different data sets are identified by this method. There are many strategies for identifying patterns, paradigms, and relationships in data sets. Additionally, developing hypotheses about these connections that can be applied to emerge previously unknown data. Because AI is very successful in predicting and categorizing, even more so in breast cancer clinical analysis, and its use in the clinical field is growing rapidly (L. K. Singh et al., 2023). In biomedical research, it is also widely used. After lung cell death, breast cancer is the second most common reason for mortality in women (Faramarzi et al., 2021). As a result, it is critical to detect breast cancer at an early stage. By separating facts from information that suggests a disease, one can build expectations about the disease. The review used a careful examination of AI tactics to improve the accuracy of breast cancer rate estimates.

Scientists developed a clever technique to identify malignant breast growths using a machine learning classifier (Omondiagbe et al., 2019). A machine learning model was created to differentiate between benign and malignant breasts by leveraging the Wisconsin Diagnostic data set (Sengar et al., 2020). To convey the ethics of ML and its prospects, numerous studies have been conducted to distinguish exemplary scalable approaches and conventional ML characterization processes. Results show that ML strategies, which are the result of developing and improving AI techniques as well as the growing volume and complexity of information, have the most prominent unwavering quality characteristics (Abdulhay et al., 2018). A group technique is used to combine several models in the demonstrated study so that the expected precision of each classifier can be compared across different types of item classes. This method combines SVM, NB, and J48 with the democratic classifier methodology to achieve a precision of 97.13, which is higher than any separate classifier (Kumar et al., 2017). Several studies have been conducted in classifying breast cancer disease using models (ML) showing good results. But how standardscaler can increase the accuracy of using machine learning algorithms to predict breast cancer. Therefore, it is necessary to conduct a continuous study in predicting breast cancer, so that it can help medical personnel to take further action and appropriate treatment.

2. Literature Review

This section reviews earlier research in the topic of data classification for breast cancer. A portion of these publications are devoted to classification schemes. The findings of earlier research will be first explained in the following.

Research conducted by Atban et al. (2023) that a publicly available benchmark dataset, BreakHis, has been used for experimental investigation of the suggested method. Experimental results show that the recommended strategy uses Support Vector Machine (SVM) with Gaussian and radial-based functions (RBF) to achieve an F-score of 97.75% for features derived from ResNet18-EO.

Botlagunta et al. (2023) in her research, removing outliers from blood profile data significantly improves the accuracy of machine learning models. With an AUC of 0.87, the Decision Tree (DT) classifier demonstrated 83% accuracy. Next, they used Flask to apply a DT classifier to build a web application for reliable diagnosis of MBC patients. All things considered, they concluded that ML models built on blood profile data could help doctors select MBC patients who require intensive treatment to improve overall survival rates.

Another study by Egwom et al. (2022) describes a classification model for breast cancer using ML. For feature classification and extraction, SVM and linear discriminant analysis (LDA), respectively, are used. The study had better results, with 99.2% accuracy, 98.0% recall, and 98.0% precision on the WBCD data set, compared to 79.5% accuracy, 76.0% recall, and 59.0% precision on the WPBC data set. When LDA is used and median is used to calculate missing values, SVM classifiers work better when handling classification issues.

Bayrak et al. (2019) used two widely used machine learning algorithms to classify the Wisconsin Breast Cancer (Native) dataset. Accuracy, precision, recall and Area ROC scores are used to compare the classification performance of these techniques with each other. The Support Vector Machine approach provides the best results with the highest accuracy.

Yadavendra & Chand (2020) used various ML methods in this work, categorizing breast cancer tumors and assessing the effectiveness of several classifiers. For the classification of breast

cancer tumors, the Xception technique performs better than any alternative method in terms of precision, memory, and F1 scores. Assiri et al. (2020) in his research, ML classifiers were used, namely ensemble classification using voting mechanisms, simple LR learning, SVM learning with stochastic gradient decrease optimization, and multilayer perceptron networks were used. Comparing the performance of the hard voting mechanism (majority-based election) with the WBCD's advanced algorithm, the hard voting mechanism performed better with 99.42%.

Ara et al. (2021) in her research, tumors were divided into benign and malignant categories using machines learning. To select the most accurate approach, each method must have a calculation and comparison of accuracy. The investigation found that the SVM and RFperformed with 96.5% accuracy better than other classifiers. This classifier can be used to develop automated diagnostic tools for the early diagnosis of breast cancer. Bayrak et al. (2019) in his study, The Wisconsin Breast Cancer Dataset (WBCD) was classified using two popular ML methods. The classification performance of these approaches was contrasted using accuracy, precision, memory, and ROC Area values. Performance is optimal when using SVM method, which offers the maximum accuracy.

Wu & Hicks (2021) evaluated four different classification methods to train a model to characterize two types of breast cancer. Compared to other ML algorithms that have been evaluated, the supporting vector engine is able to classify lung cancer as triple negative or non-triple negative, as well as having a more favorable classification threshold than the other three algorithms. Jabbar (2021) in his research, to use ensemble learning to solve categorizing breast cancer data problem. The new strategy goes beyond existing methodologies, according to experimental results, and records an astonishing accuracy of 97% when classifying breast cancer data.

Zhang et al. (2022) in his research, in identifying normal cells from breast cancer and predicting breast cancer subtypes, They try to streamline this process by leveraging Raman spectroscopy and ML approaches. Principal component analysis (PCA)-discriminant function analysis (DFA) and SVM PCA are two of the many machine learning techniques used to deal with data. Breast cancer cell lines that have been cultured are used to obtain Raman spectra. These two algorithms have an accuracy rate of more than 97% in the ability to distinguish between breast cancer and healthy cells, and more than 92% in the ability to classify breast cancer subtypes.

Laghmati et al. (2020) in her research, the machine learning technique was tested and then trained using WBCD. Features loaded from the data set are implemented into the model so that when feature selection can use Environmental Component Analysis (NCA), which reduces the number of features and model complexity. The best predictive specificity is the 9S. S6% for Binary SVM models, and maximum predictive sensitivity up to one for KNN and Adaboost models. The highest prediction accuracy was 99.12% for the KNN model.

3. Research Methods

By using ML algorithms and collecting data, breast cancer can be classified using machine learning. Figure 1 general steps of using machine learning to classify breast cancer.

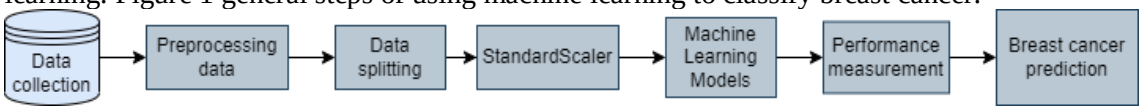

Fig. 1. Research Methods

The research steps consisted of collecting breast cancer patient datasets, preprocessing data, data separation, data transformation using StandardScaler, machine learning models, performance measurement, and breast cancer prediction using confusion matrix. Each process has its own tasks to achieve the desired goals. In this research, we added a StandardScaler process for the reason of increasing performance measurement accuracy.

3.1 Data Collection

The practice of gathering information or data from diverse sources for analysis, investigation, or decision-making is known as data collection (Yang et al., 2022). The first step in comprehending a certain occurrence or issue is gathering data. The data collection process in this research can be taken for free from the Kaggle site.

3.2 Preprocessing Data

Before being used by machine learning algorithms, data must be processed. This procedure can involve eliminating unnecessary values, standardizing data, and selecting features (features most related to categorization) (Maharana et al., 2022). Determine the missing values in the data set, and then decide how to treat them. Some alternatives include deleting rows or columns containing mASissing values [Shahidul Islam Khan], or using more complex methods such as interpolation (Wang et al., 2020). It is important to encode category variables into numerical values for analysis or modeling if the data set contains such variables (Budholiya et al., 2022). Depending on the type of data being processed and the algorithm used, this can be achieved using methods such as label coding.

3.3 Data Splitting

It is necessary to separate the data into two categories: training data and testing data (Medin & Smith, 1981). The model was trained using training data, and its effectiveness was evaluated using test data (Bao et al., 2019). A technique called "random separation" involves creating random subsets from the data set (Mamdouh Farghaly et al., 2023). To illustrate, you can divide the dataset into 20% for testing and 80% for training. If your data set is large enough and accurately represents the population, this random separation is helpful.

3.4 StandardScaler

One of the most widely used techniques for data pre-processing or data normalization in machine learning is StandardScaler. Each numerical feature (column) in the data set must be changed by StandardScaler so that it has a mean of zero and a standard deviation of one (G et al., 2022). Utilizing StandardScaler has the advantage of maintaining a consistent scale between numerical characteristics in the data set. When using machine learning techniques, it can help be sensitive to data size (de Amorim et al., 2023).

3.5 Model Machine Learning

The problem to be solved at this time, we try to take advantage of some of the well-known methods of ML. Including KNN, then SVM, there is also RF, the other well-known GB, LR, and also GNB.

3.5.1 K-Nearest Neighbor (KNN)

Concerning classification and regression problems, ML technique KNN is employed (Ertuğrul & Tağluk, 2017). Nearest neighbor-based learning algorithms include instance-based KNN algorithms. Finding the nearest neighbor K from a new data point in the feature space is a basic principle of KNN. KNN presupposes that data with related features will have associated labels. As a result, KNN considered the label of the nearest neighbor when making predictions on the new data and chose the majority label as the prediction (Z. Zhang et al., 2018).

3.5.2 Support Vector Machine (SVM)

Regarding regression and prediction issues, SVM is a frequently used ML technique (Ali et al., 2021). A hyperplane (dividing plane) in a feature space is constructed using SVM learning techniques to maximize the distance between samples belonging to different classes. Finding a hyperplane that can distinguish the two classes by the largest margin is the basic idea of SVM. The margin is the separation between the nearest sample in each class and the hyperplane. A hyperplane with a maximum-margin hyperplane, which SVM sought, is known as a maximum margin hyperplane(Rizwan et al., 2021).

3.5.3 Random Forest (RF)

RF is an ensemble learning strategy that deserves to be utilized in classification and regression. A "forest" is what is created when several separate decision trees are combined (Vos et al., 2017). A random subset of the training data and a random subset of the feature set were used in the construction of every tree in RF (Svetnik et al., 2003).

3.5.4 Gradient Boosting (GB)

Gradient Boosting is an ensemble learning approach or strategy that combines several weak or simple prediction models to create strong predictive models. Regression and classification problems are often addressed using this technique. Gradient Boosting involves creating predictive models sequentially, with each subsequent model concentrating on correcting errors caused by the previous model. By focusing on the gradient (subtraction) of the loss function, which is used to calculate the difference between the model's prediction and the actual value of the training data, this process is carried out. New models are introduced into the ensemble each iteration, and are selected by optimizing the gradient of the loss function compared to the current error. In order for the model ensemble as a whole to become more adept at forecasting the right results, each subsequent model strives to correct shortcomings that the previous model did not address (Licheng Zhang & Zhan, 2017).

3.5.5 Logistic Regression (LR)

One machine learning technique used for categorization problems is logistic regression. Although the word "regression" is in its name, LR is actually used to estimate the likelihood of binary outcomes (e.g., class "1" or "0") based on input variables or features. Logistic or sigmoid functions are used by logistic regression algorithms to represent the relationship between input data (in the form of real numbers) and binary output variables. The output is converted by the sigmoid function into a number between 0 and 1, which represents the probability of a successful outcome. A higher probability of a positive outcome is indicated by a value close to 1, while a greater probability of a negative event is indicated by a value close to 0 (Tu, 1996).

3.5.6 Gaussian Naive Bayes (GNB)

The Naive Bayes family of algorithms includes a classification algorithm known as Gaussian NB (Naive Bayes). For classifications based on Bayes' theorem, this approach is often used in machine learning (G. Singh et al., 2019). From the premise that the features used for classification are normally distributed (or Gaussian), Gaussian NB is based. Based on the possible features seen in the training data, this algorithm generates the probability of the class (Shiri Harzevili & Alizadeh, 2018).

3.6 Performance Measurement

Various metrics of evaluation that are widely used to measure model performance in ML, particularly in the context of classification. Some significant performance metrics are as follows:

3.6.1 Accuracy

The easiest and most popular metric to measure how well a model can perform accurate categorization is accuracy. By dividing the number of accurate predictions by the entire amount of data, accuracy can be obtained in this way. However, when the data is uneven or the class is relatively sparse, accuracy is not necessarily the most revealing metric. Determining accuracy can be done with Equation (1).

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

The correct positive and negative numbers are TP and TN, respectively. False positives and false negatives are measured by the letters FP and FN, respectively.

3.6.2 Precision and Recall

Precision and recall can be used to measure performance in detecting positive classes. Precision measures the accuracy of the model's positive predictions, whereas recall assesses how well the model can locate each instance of a genuine positive class. Calculation of precision is positive class occurrences total number divided by correct positive predictions number. Whila recall calculation is positive predictions number divided by correct positive predictions number. Determining precision can be done with Equation (2), and recall with Equation (3).

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

3.6.3 F1-Score

F1 scores combine memory and precision into a single number. F1 scores result in a balanced average of precision and memory between the two, resulting in a misaligned average. The F1 score is determined by multiplying the precision and recall numbers twice and dividing the result by the total number of precision and recall. Determining precision can be done with Equation (4).

$$F1\text{-}Score = (2 \times Precision \times Recall) / (Precision + Recall)$$

4. Results and Discussions

After the steps in the study are carried out which aims to classify benign and malignant breast cancer, then here we describe the results of the methods that have been done.

4.1 Data Collection

The breast lump dataset underwent a biopsy to classify it as malignant (cancerous) or benign (not cancerous). Digital images of fine needle aspiration biopsy slides are used to computationally extract features. The size, shape, and regularity of features correspond to the cell nucleus. For a total of 30 features, the mean, standard deviation, and worst values of each of the 10 nuclear parameters are presented in Table 1 (*Breast Cancer Wisconsin Diagnostic Dataset | Kaggle*, n.d.).

Table 1 - Breast Cancer Dataset

	radius mean	Texture mean	Perimeter mean	Area mean	…	Texture worst	Area worst	Concavity mean	Symmetry worst	Y
0	13.540	14.36	87.46	566.3	…	19.26	711.2	0.06664	0.2977	B
1	13.080	15.71	85.63	520.0	…	20.49	630.5	0.04568	0.3184	B
2	9.504	12.44	60.34	273.9	…	15.66	314.9	0.02956	0.2450	B
3	13.030	18.42	82.61	523.8	…	22.81	545.9	0.02562	0.1987	B
4	8.196	16.84	51.71	201.9	…	21.96	242.2	0.01588	0.3105	B
…	…	…	…	…	…	…	…	…	…	…
564	20.920	25.09	143.00	1347.0	…	29.41	1819.0	0.31740	0.2929	M
565	21.560	22.39	142.00	1479.0	…	26.40	2027.0	0.24390	0.2060	M
566	20.130	28.25	131.20	1261.0	…	38.25	1731.0	0.14400	0.2572	M
567	16.600	28.08	108.30	858.1	…	34.12	1124.0	0.09251	0.2218	M
568	20.600	29.33	140.10	1265.0	…	39.42	1821.0	0.35140	0.4087	M

The dataset presented in Table 1 consists of 30 features to predict breast cancer with benign and malignant values. A total of 569 data will be trained and tested through 30 features consisting of; radius is the radius of the nucleus (the average distance from the center to points on the circumference), texture is the texture of the nucleus (standard deviation of grayscale values), perimeter is the perimeter of the nucleus, area is the area of the nucleus, smoothness is the smoothness of the nucleus (local variation in radius length), concavity is the compactness of the nucleus (perimeter^2/area - 1), concave point is the concave of the nucleus (severity of the concave part of the contour), symmetry is the symmetry of the nucleus, and the fractal dimension is the fractal dimension of the nucleus ("approximate coastline" -1). The Y feature as a target is a two-level factor that indicates whether a mass is malignant ("M") or benign ("B").

4.2 Preprocessing Data

At this stage, unnecessary data cleaning is carried out. As seen in Figure 2, the Unnamed variable is not needed. These variables are omitted so as not to interfere in the classification process. The process of replacing the target variable Y is also carried out, so that it is clearly visible the variable that is the target in this dataset. Then the most important thing is to convert the target data which is still categorical into numerical variables, so that the ML classification process can run well. The results of preprocessing can be seen in Figure 2.

ter_worst	x.area_worst	x.smoothness_worst	x.compactness_worst	x.concavity_worst	x.concave_pts_worst	x.symmetry_worst	x.fractal_dim_worst	BreastCancer
99.70	711.2	0.14400	0.17730	0.23900	0.12880	0.2977	0.07259	0
96.09	630.5	0.13120	0.27760	0.18900	0.07283	0.3184	0.08183	0
65.13	314.9	0.13240	0.11480	0.08867	0.06227	0.2450	0.07773	0
84.46	545.9	0.09701	0.04619	0.04833	0.05013	0.1987	0.06169	0
57.26	242.2	0.12970	0.13570	0.06880	0.02564	0.3105	0.07409	0

Fig. 2. Dataset after preprocessing

Figure 2 shows the condition of the dataset that has been preprocessed. The target variable Y has been changed to BreastCancer. While the data on the target variable has been converted into numerical data. The value of "B" categorized as benign is changed to 0, and the value of "M" categorized as malignant is changed to 1. After this, the dataset can be processed further at the data separation stage. Before that we show the variable distribution of breast cancer in Figure 3.

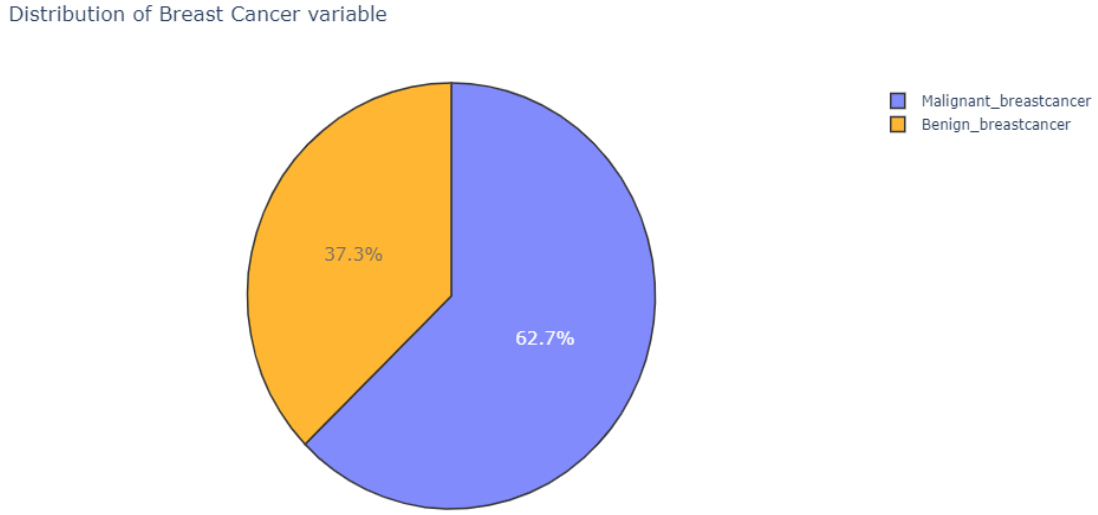

Fig. 3. Variable Distribution of Breast Cancer

In Figure 3, you can see the difference in the presentation of the target variable value. The number of breast cancer scores was 62.7%, higher than the benign breast cancer score of 37.3%.

4.3 Data Splitting

The data separation stage is carried out by separating the training data from the test data. Training data was taken as much as 75% of the total dataset, and test data was taken as much as 25% of the total dataset. So that in the next process, the classification process can be carried out by several models in ML.

4.4 StandardScaler

The scikit-learn (sklearn) library in python contains the StandardScaler implementation. Figure 4 is the form of the script we used when StandardScaler was implemented after data splitting was done.

```
std = StandardScaler()
X_train = std.fit_transform(X_train)
X_test = std.transform(X_test)
```

Fig. 4. Standardscaler Implementation

StandardScaler is used to transform both training datasets and test datasets. This is implemented so that performance performed using ML results in better performance.

4.5 Performance Measurement

Accuracy, precision, F1-score, and recall are measured as a function of various ML methods performance. The tests conducted on each model yielded the following results, which are presented below.

Table 2 – ML Algorithm Performance Measurement Results

Algorithm	Benign			Malignant			Accuracy
	Prec.	Recall	F1-Score	Prec.	Recall	F1-Score	
KNN	0.96	0.99	0.97	0.98	0.93	0.95	0.97
SVM	0.98	1.00	0.99	1.00	0.96	0.98	0.99
RF	0.97	1.00	0.98	1.00	0.95	0.97	0.98
GB	0.96	1.00	0.98	1.00	0.93	0.96	0.97
LR	0.99	1.00	0.99	1.00	0.98	0.99	0.99
GNB	0.96	1.00	0.98	1.00	0.93	0.96	0.97

Table 2 is the performance measurement results of several ML algorithms. The KNN algorithm shows precison results of 0.96 benign, and 0.98 malignant. While the recall result of 0.99 is benign, and 0.93 is malignant. The F1-Score shows 0.97 benign, and 0.95 malignant. The accuracy obtained from the KNN algorithm is 0.97. The SVM algorithm shows a precison result of 0.98 benign, and 1.00 malignant. While the recall results of 1.00 are benign, and 0.96 are malignant. The F1-Score shows 0.99 benign, and 0.98 malignant. The accuracy obtained from the SVM algorithm is 0.99. The RF algorithm shows a precison result of 0.97 benign, and 1.00 malignant. While the recall results of 1.00 are benign, and 0.95 are malignant. The F1-Score shows 0.98 benign, and 0.97 malignant. The accuracy obtained from the RF algorithm is 0.98. The GB algorithm shows a precison result of 0.96 benign, and 1.00 malignant. While the recall results of 1.00 are benign, and 0.93 are malignant. The F1-Score shows 0.98 benign, and 0.98 malignant. The accuracy obtained from the GB algorithm is 0.98. The LR algorithm shows precison results of 0.99 benign, and 1.00 malignant. While the recall results of 1.00 are benign, and 0.98 are malignant. The F1-Score results show 0.99 benign, and 0.99 malignant. The accuracy obtained from the LR algorithm is 0.99. The GNB algorithm shows a precison result of 0.96 benign, and 1.00 malignant. While the recall results of 1.00 are benign, and 0.93 are malignant. The F1-Score results show 0.98 benign, and 0.96 malignant. The accuracy obtained from the RF algorithm is 0.97.

4.6 Breast Cancer Prediction

Confusion Matrix is used to show the results of the prediction algorithm used in this study. Here we show the prediction results of each algorithm.

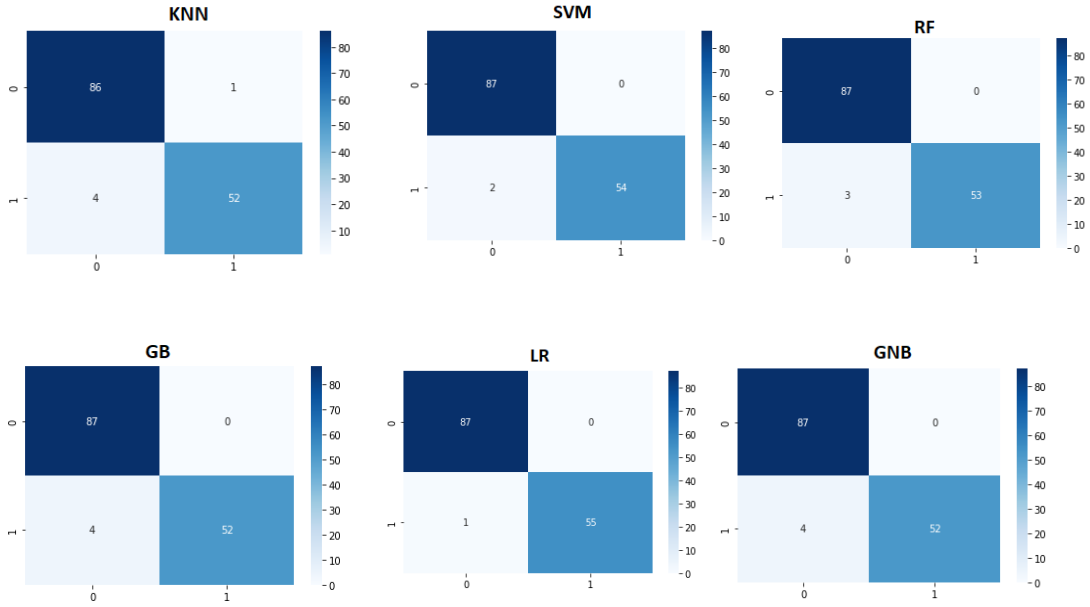

Fig. 5. ML Algorithm Confusion Matrix Results

Figure 5 is the result of the confusion matrix of the ML algorithm. The KNN algorithm displays 138 total correct predictions and 5 total incorrect ones. The SVM algorithm displays a total of 141 accurate predictions and 2 inaccurate ones. According to the RF algorithm, there were 140 total right predictions and three incorrect ones. According to the GB algorithm, there were 139 correct predictions overall and four wrong ones. According to the LR algorithm, there were 142 correct predictions overall, and one incorrect prediction. Additionally, the GNB algorithm indicates that there were up to 139 correct predictions overall, whereas there were up to four incorrect ones. Accuracy-based comparison of the suggested approach with a few prior works can be seen in Table 3.

Table 3 – Accuracy-Based Comparison of The Suggested Approach With A Few Prior Works

Reference	Year	Algorithm	Accuracy
Jabbar	2021	Bayesian	97%
Wu & Hicks	2021	SVM	90%

Ara et al.	2021	Random Forest, SVM	96,5%
Bayrak et al.	2019	SVM	95,42%
Hazra et al.	2020	ANN, Decision Tree	98,55%
Khandezamin et al.	2020	Logistic Regression	99,1%
Egwom et al.	2022	LDA-SVM	99,2%
Atban et al.	2023	SVM, Gaussian	97,73%
Botlagunta et al.	2023	Decision Tree (DT) classifier	83%
Bhanushali et al.	2023	Random Forest	93,8%
Amin et al.	2023	SVM	98%
Manikandan et al.	2023	Decision Tree	98%
Safdar et al.	2022	K-nearest neighbor	97,7%
Rabiei et al.	2022	Random Forest	80%
Monirujjaman Khan et al.	2022	Logistic Regression	98%
Proposed Method	2023	Logistic Regression	99,3%

The results of the accuracy values shown in table 3 show that the accuracy results of the proposed method have increased compared to previous research. Sequentially, the highest accuracy value shown is 99.3% obtained from the method proposed in this research, then 99.2% in the research of Egwom et al., (2022), and 99,1% by the research of Khandezamin et al., (2020).

5. Conclusion

The breast cancer diagnosis dataset that we have tested, yielded an excellent classification. Accuracy is very satisfactory, as is precision, recall, and the F1-Score score is also very satisfactory, showing how reliable the classifier is. A total of 6 algorithms used on average achieved accuracy above 96%. The use of SS in this study has an impact on performance results and predictions using ML algorithms. It can be seen that the recall results of benign diagnosis and the precision results of malignant diagnosis on average almost reach 100%. While the highest accuracy in this study was obtained from the LR algorithm, which is 99,3%. This proves that testing breast cancer diagnosis datasets using ML produces excellent performance and prediction, with the help of SS to optimize in training and testing data transformation. So that this testing can help the medical party in the follow-up of patients infected with breast cancer.

References

Abdulhay, E., Mohammed, M. A., Ibrahim, D. A., Arunkumar, N., & Venkatraman, V. (2018). Computer Aided Solution for Automatic Segmenting and Measurements of Blood Leucocytes Using Static Microscope Images. *Journal of Medical Systems*, *42*(4), 1–12. https://doi.org/10.1007/S10916-018-0912-Y/METRICS

Ali, L., Wajahat, I., Amiri Golilarz, N., Keshtkar, F., & Bukhari, S. A. C. (2021). LDA–GA–SVM: improved hepatocellular carcinoma prediction through dimensionality reduction and genetically optimized support vector machine. *Neural Computing and Applications*, *33*(7), 2783–2792. https://doi.org/10.1007/S00521-020-05157-2/METRICS

Amin, S. A., Al Shanabari, H., Iqbal, R., & Karyotis, C. (2023). An Intelligent Framework for Automatic Breast Cancer Classification Using Novel Feature Extraction and Machine Learning Techniques. *Journal of Signal Processing Systems*, *95*(2–3), 293–303. https://doi.org/10.1007/S11265-022-01753-8/METRICS

Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018). Breast cancer classification using machine learning. *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, EBBT 2018*, 1–4. https://doi.org/10.1109/EBBT.2018.8391453

Ara, S., Das, A., & Dey, A. (2021). Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms. *2021 International Conference on Artificial Intelligence, ICAI 2021*, 97–101. https://doi.org/10.1109/ICAI52203.2021.9445249

Assiri, A. S., Nazir, S., & Velastin, S. A. (2020). Breast Tumor Classification Using an Ensemble Machine Learning Method. *Journal of Imaging 2020, Vol. 6, Page 39*, *6*(6), 39. https://doi.org/10.3390/JIMAGING6060039

Atban, F., Ekinci, E., & Garip, Z. (2023). Traditional machine learning algorithms for breast cancer image classification with optimized deep features. *Biomedical Signal Processing and Control*, *81*, 104534. https://doi.org/10.1016/J.BSPC.2022.104534

Bao, S., He, H., Wang, F., Wu, H., & Wang, H. (2019). PLATO: Pre-trained Dialogue Generation

Model with Discrete Latent Variable. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 85–96. https://doi.org/10.18653/v1/2020.acl-main.9

Bayrak, E. A., Kirci, P., & Ensari, T. (2019). Comparison of machine learning methods for breast cancer diagnosis. *2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science, EBBT 2019*. https://doi.org/10.1109/EBBT.2019.8741990

Bhanushali, A., Sivagnanam, K., Singh, K., Mittapally, B. K., Reddi, L. T., & Bhanushali, P. (2023). Analysis of Breast Cancer Prediction Using Multiple Machine Learning Methodologies. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(3), 1077–1084. https://ijisae.org/index.php/IJISAE/article/view/3367

Botlagunta, M., Botlagunta, M. D., Myneni, M. B., Lakshmi, D., Nayyar, A., Gullapalli, J. S., & Shah, M. A. (2023). Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. *Scientific Reports 2023 13:1*, *13*(1), 1–17. https://doi.org/10.1038/s41598-023-27548-w

Breast Cancer Wisconsin Diagnostic Dataset | Kaggle. (n.d.). Retrieved June 1, 2023, from https://www.kaggle.com/datasets/utkarshx27/breast-cancer-wisconsin-diagnostic-dataset

Budholiya, K., Shrivastava, S. K., & Sharma, V. (2022). An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University - Computer and Information Sciences*, *34*(7), 4514–4523. https://doi.org/10.1016/J.JKSUCI.2020.10.013

Chakraborty, S., Aich, S., & Kim, H. C. (2019). A Secure Healthcare System Design Framework using Blockchain Technology. *International Conference on Advanced Communication Technology, ICACT*, *2019-February*, 260–264. https://doi.org/10.23919/ICACT.2019.8701983

Cios, K. J., & William Moore, G. (2002). Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, *26*(1–2), 1–24. https://doi.org/10.1016/S0933-3657(02)00049-0

de Amorim, L. B. V., Cavalcanti, G. D. C., & Cruz, R. M. O. (2023). The choice of scaling technique matters for classification performance. *Applied Soft Computing*, *133*, 109924. https://doi.org/10.1016/J.ASOC.2022.109924

Egwom, O. J., Hassan, M., Tanimu, J. J., Hamada, M., & Ogar, O. M. (2022a). An LDA–SVM Machine Learning Model for Breast Cancer Classification. *BioMedInformatics 2022, Vol. 2, Pages 345-358*, *2*(3), 345–358. https://doi.org/10.3390/BIOMEDINFORMATICS2030022

Egwom, O. J., Hassan, M., Tanimu, J. J., Hamada, M., & Ogar, O. M. (2022b). An LDA–SVM Machine Learning Model for Breast Cancer Classification. *BioMedInformatics 2022, Vol. 2, Pages 345-358*, *2*(3), 345–358. https://doi.org/10.3390/BIOMEDINFORMATICS2030022

Ertuğrul, Ö. F., & Tağluk, M. E. (2017). A novel version of k nearest neighbor: Dependent nearest neighbor. *Applied Soft Computing*, *55*, 480–490. https://doi.org/10.1016/J.ASOC.2017.02.020

Fadilah, D., Putri, A., Putu, L., & Yuliastuti, S. (2022). Effect of health education using demonstration media for breast self-examination motivation for women in preventing breast cancer. *Jurnal Pijar Mipa*, *17*(5), 679–682. https://doi.org/10.29303/JPM.V17I5.3993

Faramarzi, A., Jahromi, M. G., Jalilian, N., Golestan Jahromi, M., & Ashourzadeh, S. (2021). Metastatic and pathophysiological characteristics of breast cancer with emphasis on hereditary factors Improving Human Sperm Culture Medium View project Metastatic and pathophysiological characteristics of breast cancer with emphasis on hereditary factors. *Central Asian Journal of Medical and Pharmaceutical Sciences Innovation*, *3*, 104–113. https://doi.org/10.22034/CAJMPSI.2021.03.01

G, T. R., Bhattacharya, S., Maddikunta, P. K. R., Hakak, S., Khan, W. Z., Bashir, A. K., Jolfaei, A., & Tariq, U. (2022). Antlion re-sampling based deep neural network model for classification of imbalanced multimodal stroke dataset. *Multimedia Tools and Applications*, *81*(29), 41429–41453. https://doi.org/10.1007/S11042-020-09988-Y/METRICS

Hazra, R., Banerjee, M., & Badia, L. (2020). Machine Learning for Breast Cancer Classification with ANN and Decision Tree. *11th Annual IEEE Information Technology, Electronics and Mobile Communication Conference, IEMCON 2020*, 522–527. https://doi.org/10.1109/IEMCON51383.2020.9284936

Hughes, D. T., Reyes-Gastelum, D., Ward, K. C., Hamilton, A. S., & Haymart, M. R. (2022). Barriers to the use of active surveillance for thyroid cancer: Results of a physician survey. *Annals of Surgery*, *276*(1), e40. https://doi.org/10.1097/SLA.0000000000004417

Jabbar, M. A. (2021). Breast Cancer Data Classification Using Ensemble Machine Learning. *Engineering and Applied Science Research*, *48*(1), 65–72. https://doi.org/10.14456/easr.2021.8

Khandezamin, Z., Naderan, M., & Rashti, M. J. (2020). Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier. *Journal of Biomedical Informatics*, *111*, 103591. https://doi.org/10.1016/J.JBI.2020.103591

Klein, E. A., Richards, D., Cohn, A., Tummala, M., Lapham, R., Cosgrove, D., Chung, G., Clement, J., Gao, J., Hunkapiller, N., Jamshidi, A., Kurtzman, K. N., Seiden, M. V., Swanton, C., & Liu, M. C. (2021). Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Annals of Oncology*, *32*(9), 1167–1177. https://doi.org/10.1016/J.ANNONC.2021.05.806

Kumar, U. K., Nikhil, M. B. S., & Sumangali, K. (2017). Prediction of breast cancer using voting classifier technique. *2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, ICSTM 2017 - Proceedings*, 108–114. https://doi.org/10.1109/ICSTM.2017.8089135

Laghmati, S., Cherradi, B., Tmiri, A., Daanouni, O., & Hamida, S. (2020). Classification of Patients with Breast Cancer using Neighbourhood Component Analysis and Supervised Machine Learning Techniques. *3rd International Conference on Advanced Communication Technologies and Networking, CommNet 2020*. https://doi.org/10.1109/COMMNET49926.2020.9199633

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1), 91–99. https://doi.org/10.1016/J.GLTP.2022.04.020

Mamdouh Farghaly, H., Shams, M. Y., & Abd El-Hafeez, T. (2023). Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt. *Knowledge and Information Systems*, *65*(6), 2595–2617. https://doi.org/10.1007/S10115-023-01851-4/TABLES/7

Manikandan, P., Durga, U., & Ponnuraja, C. (2023). An integrative machine learning framework for classifying SEER breast cancer. *Scientific Reports 2023 13:1*, *13*(1), 1–12. https://doi.org/10.1038/s41598-023-32029-1

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(4), 241–253. https://doi.org/10.1037/0278-7393.7.4.241

Mekha, P., & Teeyasuksaet, N. (2019). Deep learning algorithms for predicting breast cancer based on tumor cells. *ECTI DAMT-NCON 2019 - 4th International Conference on Digital Arts, Media and Technology and 2nd ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering*, 343–346. https://doi.org/10.1109/ECTI-NCON.2019.8692297

Monirujjaman Khan, M., Islam, S., Sarkar, S., Ayaz, F. I., Ananda, M. K., Tazin, T., Albraikan, A. A., & Almalki, F. A. (2022). Machine Learning Based Comparative Analysis for Breast Cancer Prediction. *Journal of Healthcare Engineering*, *2022*. https://doi.org/10.1155/2022/4365855

Nozomi, I., Aldi, F., & Sentosa, R. B. (2022). Views on Deep Learning for Medical Image Diagnosis. *Journal of Applied Engineering and Technological Science (JAETS)*, *4*(1), 547–553. https://doi.org/10.37385/JAETS.V4I1.1367

Omondiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019). Machine Learning Classification Techniques for Breast Cancer Diagnosis. *IOP Conference Series: Materials Science and Engineering*, *495*(1), 012033. https://doi.org/10.1088/1757-899X/495/1/012033

Rabiei, R., Ayyoubzadeh, S. M., Sohrabei, S., Esmaeili, M., & Atashi, A. (2022). Prediction of Breast Cancer using Machine Learning Techniques. *ACM International Conference Proceeding Series*, *March*, 382–387. https://doi.org/10.1145/3549206.3549274

Rizwan, A., Iqbal, N., Ahmad, R., & Kim, D. H. (2021). WR-SVM Model Based on the Margin Radius Approach for Solving the Minimum Enclosing Ball Problem in Support Vector Machine Classification. *Applied Sciences 2021, Vol. 11, Page 4657*, *11*(10), 4657. https://doi.org/10.3390/APP11104657

Safdar, S., Rizwan, M., Gadekallu, T. R., Javed, A. R., Rahmani, M. K. I., Jawad, K., & Bhatia, S. (2022). Bio-Imaging-Based Machine Learning Algorithm for Breast Cancer Detection. *Diagnostics 2022, Vol. 12, Page 1134*, *12*(5), 1134. https://doi.org/10.3390/DIAGNOSTICS12051134

Sengar, P. P., Gaikwad, M. J., & Nagdive, A. S. (2020). Comparative study of machine learning algorithms for breast cancer prediction. *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, 796–801. https://doi.org/10.1109/ICSSIT48917.2020.9214267

Shiri Harzevili, N., & Alizadeh, S. H. (2018). Mixture of latent multinomial naive Bayes classifier. *Applied Soft Computing*, *69*, 516–527. https://doi.org/10.1016/J.ASOC.2018.04.020

Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification. *2019 International Conference on Automation, Computational and Technology Management, ICACTM 2019*, 593–596. https://doi.org/10.1109/ICACTM.2019.8776800

Singh, L. K., Khanna, M., & Singh, R. (2023). Artificial intelligence based medical decision support system for early and accurate breast cancer prediction. *Advances in Engineering Software*, *175*, 103338. https://doi.org/10.1016/J.ADVENGSOFT.2022.103338

Svetnik, V., Liaw, A., Tong, C., Christopher Culberson, J., Sheridan, R. P., & Feuston, B. P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1947–1958. https://doi.org/10.1021/CI034160G/SUPPL_FILE/CI034160GSI20031008_041202.ZIP

Tazin, T., Sarker, S., Gupta, P., Ayaz, F. I., Islam, S., Monirujjaman Khan, M., Bourouis, S., Idris, S. A., & Alshazly, H. (2021). A Robust and Novel Approach for Brain Tumor Classification Using Convolutional Neural Network. *Computational Intelligence and Neuroscience*, *2021*. https://doi.org/10.1155/2021/2392395

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, *49*(11), 1225–1231. https://doi.org/10.1016/S0895-4356(96)00002-9

van der Niet, A. G., & Bleakley, A. (2021). Where medical education meets artificial intelligence: 'Does technology care?' *Medical Education*, *55*(1), 30–36. https://doi.org/10.1111/MEDU.14131

Vos, T., Abajobir, A. A., Abbafati, C., Abbas, K. M., Abate, K. H., Abd-Allah, F., Abdulle, A. M., Abebo, T. A., Abera, S. F., Aboyans, V., Abu-Raddad, L. J., Ackerman, I. N., Adamu, A. A., Adetokunboh, O., Afarideh, M., Afshin, A., Agarwal, S. K., Aggarwal, R., Agrawal, A., … Murray, C. J. L. (2017). Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, *390*(10100), 1211–1259. https://doi.org/10.1016/S0140-6736(17)32154-2

Wang, Y., Wang, B., Tu, N., & Geng, J. (2020). Seismic trace interpolation for irregularly spatial sampled data using convolutional autoencoder. *Geophysics*, *85*(2), V119–V130. https://doi.org/10.1190/GEO2018-0699.1

Widiana, I. K., & Irawan, H. (2020). Clinical and Subtypes of Breast Cancer in Indonesia. *Asian Pacific Journal of Cancer Care*, *5*(4), 281–285. https://doi.org/10.31557/APJCC.2020.5.4.281-285

Wu, J., & Hicks, C. (2021). Breast Cancer Type Classification Using Machine Learning. *Journal of Personalized Medicine 2021, Vol. 11, Page 61*, *11*(2), 61.

https://doi.org/10.3390/JPM11020061

Yadavendra, & Chand, S. (2020). A comparative study of breast cancer tumor classification by classical machine learning methods and deep learning method. *Machine Vision and Applications*, *31*(6), 1–10. https://doi.org/10.1007/S00138-020-01094-1/METRICS

Yang, J., Xiu, P., Sun, L., Ying, L., & Muthu, B. (2022). Social media data analytics for business decision making system to competitive analysis. *Information Processing & Management*, *59*(1), 102751. https://doi.org/10.1016/J.IPM.2021.102751

Zhang, Licheng, & Zhan, C. (2017). Machine Learning in Rock Facies Classification: An Application of XGBoost. *Global Meeting Abstracts*, 1371–1374. https://doi.org/10.1190/IGC2017-351

Zhang, Lihao, Li, C., Peng, D., Yi, X., He, S., Liu, F., Zheng, X., Huang, W. E., Zhao, L., & Huang, X. (2022). Raman spectroscopy and machine learning for the classification of breast cancers. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *264*, 120300. https://doi.org/10.1016/J.SAA.2021.120300

Zhang, Z., Jiang, T., Li, S., & Yang, Y. (2018). Automated feature learning for nonlinear process monitoring – An approach using stacked denoising autoencoder and k-nearest neighbor rule. *Journal of Process Control*, *64*, 49–61. https://doi.org/10.1016/J.JPROCONT.2018.02.004