Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 04 Issue 02
2023

Research Article
https://doi.org/10.56979/402/2023

Breast Cancer Diagnosis Comparative Machine Learning Analysis Algorithms

Abdul Mustaan Madni[1], Awad Bin Naeem[1], Abdul Majid Soomro[3*], Biswaranjan Senapati[4], Alok Singh Chauhan[5], Fridous Ayub[2], and Khuram Shahzad[1]

[1]Department of Computer Science, National College of Business Administration & Economics, Multan, Pakistan.
[2]Department of Computer Science, Women University Swabi, Pakistan.
[3]Department of Computer Science (FSKTM), University Tun Hussein Onn Malaysia, Malaysia.
[4]Department of Computer and Data Science, Associate Professor, University of Arkansas Little Rock, Little Rock, USA.
[5]Department of Information Technology, Associate Professor, ABES Engineering College, Ghaziabad, India.
[*]Corresponding Author: Abdul Majid Soomro. Email: gi180004@siswa.uthm.edu.my

Received: October 29, 2022 **Accepted:** November 20, 2022 **Published:** March 29, 2023.

Abstract: Women have an extremely high chance of contracting breast cancer, which is often undiscovered. It is difficult to pinpoint the genesis of this ailment since it depends on several factors. In contrast, determining whether a cancer is benign or malignant requires a significant amount of effort on the part of doctors and medical professionals. Numerous assessments, such as cell size and uniformity, clump thickness, and other factors, are also taken into consideration. Additionally, this has increased the use of machine learning classifiers and their use as research tools. Using other aspects of artificial intelligence, advanced recovery rooms often depend on equipment that is qualified, while claiming that breast cancer is not. The study's objective is to assess the precision, accuracy, f1 score, specificity, recall, and error rate of machine learning tools that may be used to diagnose breast cancer. In terms of overall dataset correctness, SVM and artificial neural networks have a high accuracy of 97.45% with an error rate of 0.0154.

Keywords: Breast Cancer, Error rate, Machine learning, Data science.

1. Introduction

Breast cancer is the most often mentioned illness in the world. The order in which cancer treatments are discovered is critical. A strategy is needed to accurately detect the symptoms and match them with the known feature sets to establish the presence of malignancy. Breast cancer is classified into many types, which may assist choose the best possible therapy. The most prevalent categorization is the conditional "if the cancer is in a benign stage, less obtrusive" [1]. The frequency of therapy is utilized in the malignant stage, and the patient's survival rate is great. It is not desirable to hasten recovery at the price of potentially life-threatening side effects introduced by intensive care [2]. Furthermore, a patient with cancer is not jeopardized by the medical approach or the procedure's negative influence.

Machine learning methods are widely used in smart healthcare systems. Especially for traditional breast cancer (BC) diagnoses, a patient's medical quality is dependent on the skill of a practitioner. Nonetheless, this approach has evolved through many centuries of varied medical symptoms and proven diagnoses, albeit accuracy cannot be guaranteed [3]. With the advancement of computer technology, it is now easier to store and analyses massive amounts of data. A large quantity of data on medical patients is stored in databases. It is difficult for health practitioners to review a vast number of databases without the use of computers, especially when undertaking complicated data interrogations[4]. The integrated health care programmed is a distinct and important area. Intelligent healthcare solutions will help clinicians treat patients more precisely and with more realistic benchmarks[5]. This will assist individuals in maintaining their medical wellness in the future. For this reason, machine learning methods may be used to replace certain complex manual labor performed by physicians, such as text and voice analysis, which is used to identify emotions and their reactions in healthcare professionals. According to research, patients' emotions are the most significant to identify their health and utilize for future improved results [6].

Breast cancer seems to be the most common malignancy among women (BC). Breast cancer has a high frequency and mortality rate. According to current cancer statistics, 25% of new cases are registered, and 15% die. According to the International Health Organization, this illness causes 8.8 million deaths globally each year. Many people die as a result of cancer [7]. Cancer rates in the United States increased by 39% between 1989 and 2015. Breast cancer is the leading cause of cancer deaths in Malaysia. Women in Malaysia are at a 5% risk, whereas data reveals that women in the United States are at a 12.5% risk. Breast cancer is the most common kind of cancer in Pakistan. According to the most recent data, the top three malignancies in all age groups and both sexes were breast cancer, leukemia of the lip, and oral cavity cancer. Breast cancer claimed the lives of 40,000 individuals in 2017, with an additional 90.000 instances recorded in the same year. The most prevalent cause of breast cancer, which affects around 8 to 9 women in Pakistan, is genetic inclusion, lifestyle, and environmental impact [8].

The paper's contribution is a comparison of 5 machine learning classifiers: (ANN), (RF), (KNN), (LR), and (SVM), which are the most relevant techniques in the research community. Our study's goal is to examine precision, accuracy, f1 score, specificity, recall, and error rate.

The main goals of doing this research are:

- Finding an appropriate collection of characteristics for the specified classifiers to enhance breast cancer diagnosis.
- Identifying an appropriate dataset that may be utilized for reliable cancer diagnosis while requiring little processing overhead.
- Calculating recall, F1 score precision, and accuracy for a reliable detection procedure with fewer characteristics while maintaining acceptable level values for breast cancer diagnosis.
- Classifier analysis (as in objective 3) for quality goals with fewer features.

This work is organized as follows: Section 2 contains the literature review. Section 3 presents the methodology of the study. Section 4 discusses the results. In last, the conclusion and future work is described in section 5.

2. Literature Review

Numerous papers were given in a literature review for breast cancer diagnostics and prognoses during the previous few centuries. Previously, work on identification and analysis was done using various data mining approaches. Classification and aggregation are more popular data mining approaches, although machine learning has been described as a way of categorizing medical applications. In this section, we will show you some of the relevant research that has previously been done in the topic area [9].

The rise in the number of deaths recorded in healthcare institutions has prompted the creation of professional diagnosis support services, which aid medical professionals in their decision-making process. Many expert systems and machine learning algorithms have been created to offer supporting information based on input knowledge. Some of the major advances include the utilization of 2D and 3D imagery for diagnosis. Breast cancer is a diverse tumor with many biological behavioral subgroups as well as clinic pathological and genomic characteristics [10]. Over the last two decades, there has been a greater understanding of multistep carcinogenesis and the role of genetic alteration in breast cancer diagnosis, treatment, and prevention. This leads to advancements in breast cancer prevention, diagnostic, and treatment methods[11].

Use Wisconsin breast cancer statistics with this method to identify cancer with the highest accuracy. These classifiers are also used to detect and assess breast cancer to identify the condition at an early stage. The early diagnosis of breast cancer signs is essential for health professionals and doctors to discover cancer at an early stage and is beneficial to patients in recovery[12]. Wisconsin datasets are used to illustrate growing real machine learning algorithms for malignancy detection utilizing three classifiers in a Wisconsin diagnostic data set. In reality, all of this work has been made to automate tumor prediction with adequate accuracy [13]. Cancer has been the most often chosen subject in the last decade. In data mining, machine learning is used to diagnose cancer, but in image processing, cellular automata predict cancer utilizing CA methods and certain image processing techniques like noise detection and removal and spot detection on cancer pictures. Breast cancer, like data mining, is forecasted and classified using machine learning algorithms [14].

Machine learning may also be used to predict cancer prognosis and recurrence. We now know, thanks to biological information technologies such as genetic testing, that the terms "lung cancer," "kidney cancer," and "breast cancer" refer to hundreds of various cell mutation patterns or misalignments [15]. According to a kidney cancer article, there are no two identical leaves in the globe, just as there are no two identical tumors on the planet; another research discovered that there are no two tumors in the same tumor as a person. Cells are genetically identical. In information theory, each treatment procedure involves three notions (Data, Information, and Knowledge) [16]. Their definitions and linkages are often called into dispute. Several difficulties were necessary to solve this question. These notions are often connected according to a hierarchy based on a cognitive scale, where the low level reflects the perception of real-world situations and the high level connects to thinking. Indeed, the don-born is seen as the fundamental entity of information, and information is regarded as the fundamental entity of knowledge. Several studies investigated the DIK hierarchy (Data-Information-Knowledge) [17]. Artificial intelligence is a vast area of study that aims to create computer systems that mimic human intellect. Machine learning (ML) is an artificial intelligence (AI) discipline. It creates algorithms that help computers adapt to new issues without having to reprogrammed them. In other words, the machine learning system "learns" from data how to solve issues [18]. This is accomplished by using statistical techniques to find patterns in a collection of data without the need for human intervention. Machine learning (ML) has been routinely utilized to predict histopathology[19].

While the study revealed fresh insight into the classifier comparison, additional research is needed for this ARNCCs Detailed Interpretation. Bags and Reinforcement ARNCCs of RO-based classifiers may provide intriguing results. ARNCCs are calculated based on several rejection slots, with a single cost environment that can be applied to a broad range of cost parameters. With the addition of a fourth dimension. ARNCCs have also been assessed against a single cost setting for several rejection windows and may be used for a wide range of cost settings when the fourth dimension is included [20]. In that investigation, we proposed a technique for identifying breast masses to improve network classification efficiency. Mammographic disease and frequent usage for healthy women The categorization tests' multifunctionality and variety On DCNN, as functions[21]. The multi-function device should be selected. Where the data volume is large and there are several features. Machine learning methodologies, on the other hand, maybe explored if the amount of data is little. This study was published. The multifunctional model often outperforms the single System function or DCNN interface solely, which is reliant on DCNN. equipped for data transfer and using XGBoost in general, the layout outperformed the others in terms of ACCs. As a result, mammography multi-features are created, and the input XGBoost architecture performs better in problem regions than typical deep learning networks of categorized objects [22]. It demonstrates that the XGBoost is effective. When it comes to an inadequate amount of accessible Education materials and performance parameters, classification outperforms deep learning networks. What are the advantages of the proposed approach? [23] The main advantages are the incorporation of DCNN learned to remove performance features from Mammography images and their combination with other texture properties and instead use the XGBoost platform to achieve a better detection output given the small number of breast cancer samples and imbalanced training results. The total ACC advancement rate in Model Learning and check pattern is 9.96% and 4.10%, respectively. We want to use a system that can detect mammography by both masses and then examine the linked applications, which can be provided more comprehensive mammography; the mass characteristics are dependent on the current. Color applications GLCM and Fire, efficiency boost for breast mass definition [24].

Ada Boost was able to work properly thanks to deep learning. A breast cancer screening strategy was proposed, as well as Diagnosis Soon. The Ada Boost method was created to build an ensemble classifier for the final prediction function [25]. As a result, our suggested solution for the estimate check has more predictability and a deep-learning classifier that outperforms the other grades. Our discussion and evaluation demonstrated an exceptional ability for rapid generalization and considerably boosted the effectiveness of forecasting the result that is instantly created by neural Networking. Making the most of deep learning from the suggested deep learning interface for Convolutional Neural Network DLA-EABA also succeeded to increase programmed efficiency[26]. The learning techniques are tailored to the particular applications of a dataset center on computational thinking, and data collection will be developed for a certain model. The suggested DLAEABS technique is particularly successful in breast cancer detection. The hospital mortality rate climbs as people gains weight. The Spectacle of the proposed approach is relatively modest in comparison to several existing solutions [27].

3. Materials and Methods

This section contains the methodology of the study.

3.1. Proposed Methodology

In the proposed work (Figure 1) Propose Methodology Diagram describes the straightforward analysis. In this paper, we use machine learning and methods (ANN, KNN, SVM, RF and LR) and deep learning algorithm (ANN) to identify cancer with a small number of features.

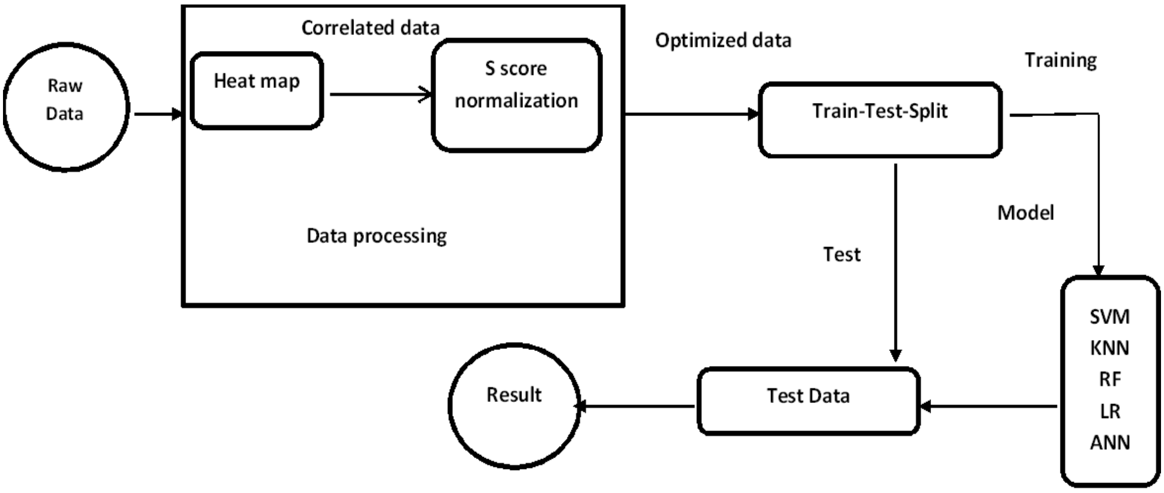


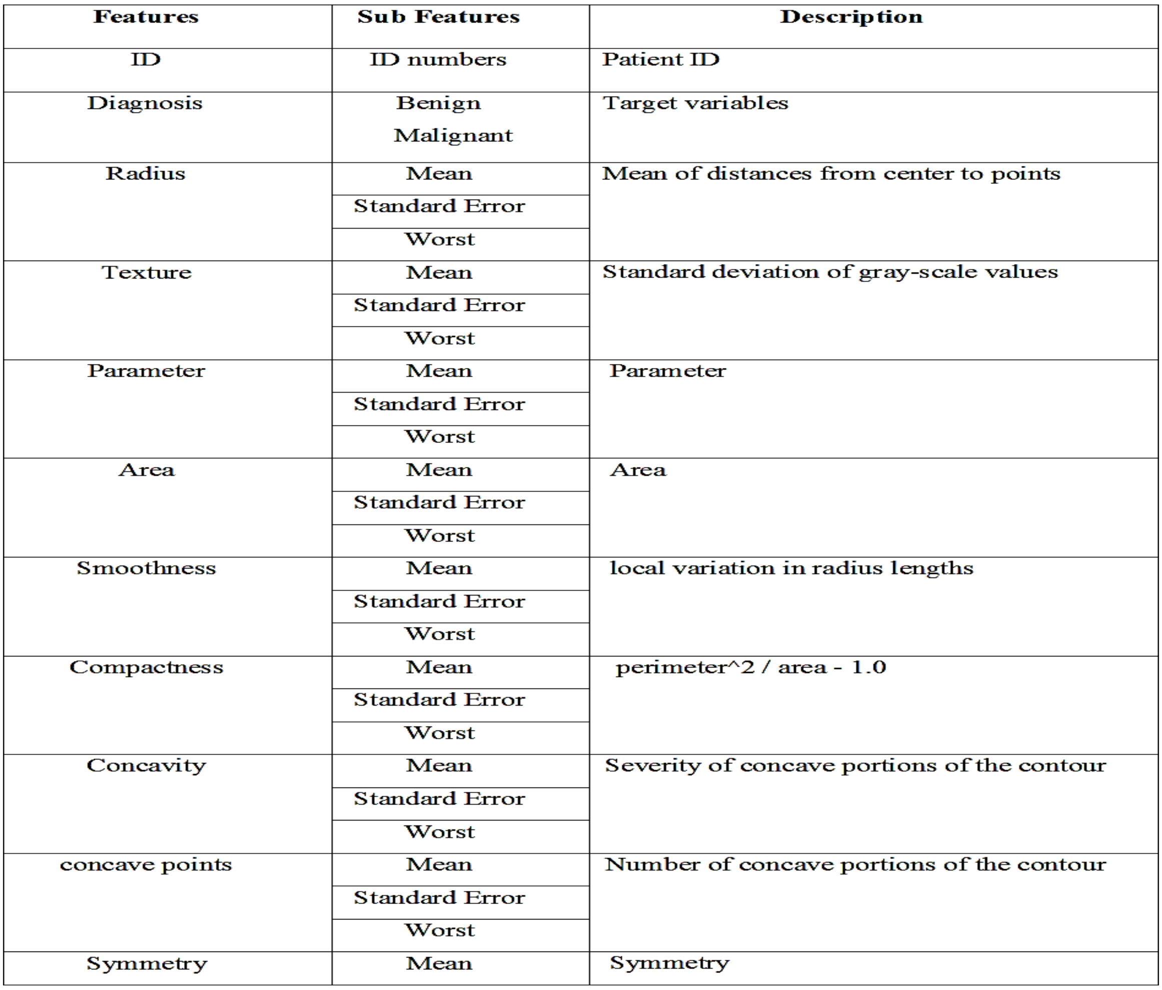
Figure 1. Propose Methodology Diagram

A supervised learning technique is employed to retain accuracy, precision, F1 score, and recall far beyond the needed standard of a breast cancer diagnosis. We obtained a dataset containing numerical values of various instances and divided the data into train and 80:20 ratio. We employ five machine learning and one deep learning classifier to test and train: Support vector machine, K-nearest neighbor, random forest, logistic regression, and artificial neural network. Then we trained the model, got the outcomes, compared them to the predicted results, and optimized them.

3.1. Data Collection

The dataset used in this paper is publicly accessible and was produced by Dr William H. Wolberg, a doctor at the University of Wisconsin Hospital in Madison, Wisconsin, USA. Dr Wolberg created the dataset using fluid tasters from women with dense breast tumor and a relatively simple-to-use digital computer tool called Xcyt, which is capable of investigating cytological aspects based on a digital scan. The sequencer uses a curve-fitting technique to compute the structures from both cells in the model, after which it evaluates the extreme value, mean value, and standard error of each feature for the picture, repeating a 30 real-valued vector.

Table 1 Dataset Description



Features	Sub Features	Description
ID	ID numbers	Patient ID
Diagnosis	Benign Malignant	Target variables
Radius	Mean	Mean of distances from center to points
	Standard Error	
	Worst	
Texture	Mean	Standard deviation of gray-scale values
	Standard Error	
	Worst	
Parameter	Mean	Parameter
	Standard Error	
	Worst	
Area	Mean	Area
	Standard Error	
	Worst	
Smoothness	Mean	local variation in radius lengths
	Standard Error	
	Worst	
Compactness	Mean	perimeter^2 / area - 1.0
	Standard Error	
	Worst	
Concavity	Mean	Severity of concave portions of the contour
	Standard Error	
	Worst	
concave points	Mean	Number of concave portions of the contour
	Standard Error	
	Worst	
Symmetry	Mean	Symmetry

(Table 1) explains the dataset description in this dataset includes three hundred fifty-seven benign cases and two hundred twelve malignant cases, with thirty-two columns in which the initial column is the patient ID and the other columns have the results of tumor detection, whether benign or malignant and surveyed by the standard deviation, mean, mean of the worst measurements of 10 features.

The dataset had no misplaced values: dataset information includes:

1. The ID of patient
2. Diagnosis (M = malignant, B = benign)

3.2 Feature Selection:

Large data evolution might be a challenging issue for scientists and engineers. Problems may be solved by deleting irrelevant and redundant data in an effective manner that is only possible via feature selection. This can speed up data processing, improve learning effectiveness, and make it easier for people to understand the learning model or data. Variable selection is often referred to as feature selection.

During the analysis, the supervised methods (Support vector machine, K-nearest neighbor, random forest, logistic regression, and artificial neural network) are employed for feature selection. Automatic selection of data characteristics that are most relevant to the challenge of predictive modelling. Dimensionality reduction and feature selection are entirely independent processes. Each strategy calls for reducing the number of features in the dataset, however, dimensionality reduction methodology is employed to create new combinations of features, whereas feature selection methodology accepts and omits characteristics discovered in the data without ever modifying those attributes.

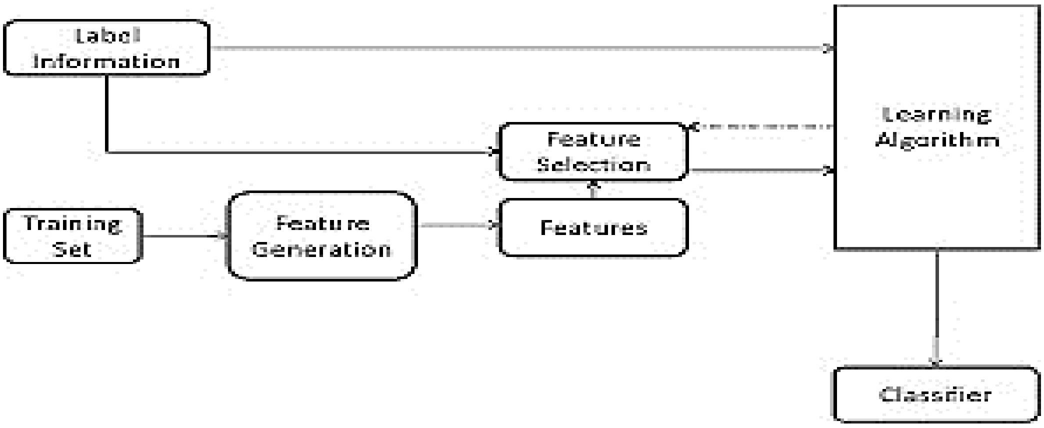


Figure 2. Feature Selection

(Figure 2) using feature selection techniques, an accurate predictive model may be produced. Choosing characteristics with assistance will result in the greatest accuracy possible while using fewer data. The feature selection approach may be used for both identification and elimination. Algorithms for feature selection often fall into one of three categories: wrapper techniques, filter methods, or embedding methods. Filter method: A statistical technique used by filter feature selection techniques to give ratings to each feature. The characteristics are either chosen to be unbroken or removed from the dataset according to a hierarchical scoring system. The techniques are often accurate and consider each characteristic separately or about the number of features.

3.2.1 Wrapper Approach:

With this method, consider group selection and alternatives as a search problem in situations where various characteristics are prepared and ready to be measured and studied with various combinations. A prediction model employs customary evaluations to rank a variety of combinations and provide pattern precision assistance. Unlike a best-first search, the hunt operation is planned and may utilize arbitrary hill climbing algorithms, heuristics like forward and backwards, to present and remove options, or an embedded technique to find the answer that improves accuracy and performance when the model is twisted. The most well-liked technique is the regularization technique and typical embedded function form. Regularization procedures, also known as penalization approaches, such as bias models with lesser complexity, provide additional restrictions on the use of predictive algorithms. We did not utilize it since it cannot be used in algorithms under any circumstances. And our goal variable is Function Selection.

3.3 Data Preprocessing

In this research, data is optimized by computing correlation using a Heat-map (to decrease the dataset) and then standardizing using Z-score normalization, as stated below. To prevent incorrect loader mapping, the data set was normalized using Eq. 1.

$$z = \frac{x- \mu}{\sigma} \qquad (1)$$

Where X is the standardizer function, is the function's defined mean value in equation 2, and σ is the standard deviation of the feature defined in equation 3. The standardization was carried out using StandardAero (). fit transform () of scikit-learn.

With mean: $$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i) \qquad (2)$$

Standard Scalar works by converting your data so that its distribution has an average value of 0 and a standard deviation of 1. When viewing the data distribution, each value in the data set will have the average value of the data subtracted and separated by the standard deviation of the entire data set.

With standard deviation: $$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)} \qquad (3)$$

3.4 Data Analysis:

For data analysis, we employed a variety of machine learning methodologies, including principal component analysis (PCA), a statistical methodology.

Use orthogonal transformation to convert a series of theoretically associated variables into a collection of linearly uncorrelated variables called PCA in which the dataset is composed of a variety of variables linked to each other it can be large or tiny by keeping the variability present inside the dataset to the greatest extent possible identification is done by renovating variables to be the set of variables to be built original variables are present in the dataset of the PCA algorithm to be utilized will be given, and the results are rather responsive. This approach is used to summaries information. Imagine a row of wine bottles on a board, with each wine described by its attributes such as color, intensity, and so on. The association rule and the heat map (static methods). It is a dimensions reduction approach with related qualities. To minimize a vast dataset and show it in a few variables. Better outcomes are obtained when data analysis tools are used to correlate characteristics. In this investigation, we employed a heat map, an ancient yet effective approach for determining the association between features. The primary function of a heat map is to discover similarities in a sample and to find correlations and variations between individual variables based on threshold values. Following the application of the heat map, the most associated characteristics were chosen, while less correlated features were removed from the dataset.

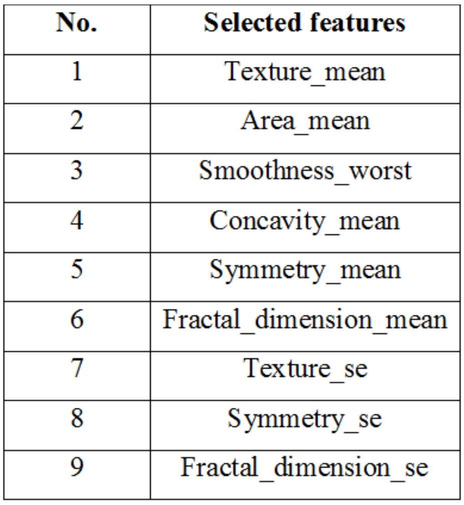
Correlation is found in two distinct threshold levels. (1) We set the threshold value to one (2) In the second example, we set the value to zero point seven.

Table 2: Selected features with thresholds value 1.0

No.	Selected Features
1	Texture_mean
2	Area_mean
3	Smoothness_mean
4	Concavity_mean
5	Symmetry_mean
6	Fractal_dimension_mean
7	Texture_se
8	Area_se
9	Smoothness_se
10	Concavity_se
11	Fractal_dimension_se
12	Texture_worst
13	Area_worst
14	Smoothness_worst
15	Concavity_worst
16	Fractal_dimension_worst

In the second scenario, we fixed a threshold value of 1. One point chose the most associated values while discarding the others. (Table 2) illustrates the 16 qualities we chose from the dataset. After analyzing the data, I believe the Concavity mean is superior to the Compactness mean, Concavity means, and concave points mean. As a result, choose an area from Radius-se, perimeter-se, and area-se. I chose the worst spot. The radius is the worst, the perimeter is the worst, and the area is the worst. Compactness worst, concavity worst, and concave points worst are my choices. Concavity se was chosen above Compactness se, concavity se, and concave points se. I chose texture to mean because texture mean and texture worst area are connected. The area means is used here.

Table 3: Selected features with thresholds value 0.7



No.	Selected features
1	Texture_mean
2	Area_mean
3	Smoothness_worst
4	Concavity_mean
5	Symmetry_mean
6	Fractal_dimension_mean
7	Texture_se
8	Symmetry_se
9	Fractal_dimension_se

In the second scenario, we fixed a threshold value of 0. Zero point seven and chose the most associated values while discarding the others. (Table 3) illustrates the 9 qualities we chose from the 16 we described earlier.

3.5 Train-Test Split:

In most situations, machine learning data is divided into two categories: testing data and training data, or into three practices: validate, check, and learn. Fits our model to the data for training. Test data is included in the actual dataset; algorithms must be trained. The model developed and learned from results was utilized for testing of data collecting outcomes to provide an impartial analysis of the final model fit to the testing dataset. Data testing is the best standard for pattern validation and is utilized once model training is completed.

The dataset is separated into training and validation groups, and the validation check may be divided into two groups. Initially, it is determined by the total number of samples in the data. Second, it examines the specific model that the user is training. To train, the majority of the models need large amounts of data. In this case, one might prepare with a variety of training sets. Some models are easy to validate and adapt, and they may minimize the size of the validation size, especially if the model includes numerous hyperparameters and user requests for validation, which include huge datasets. In this article, the proportion of splitting dataset is eighty% for training data and twenty% for the testing dataset. The training dataset has four hundred occurrences, whereas the testing dataset contains one hundred eight-nine instances. It is critical to train our algorithms in machine learning. We reserve seventy-three% of research for training and seven% for cross-validation. The cross-testing step entails dividing the data segment into complementary subsets, conducting an experiment on one subset (training groups), and assessing an experiment on the other subset (data testing). To reduce uncertainty, numerous rounds of cross-validation are performed using various divisions of several approaches. Validation tests are combined, and this round provides an estimate of the model's prediction effectiveness.

4. Results and Discussions

Following the application of a machine learning algorithm to a specific dataset, the next step is to investigate the results to determine how good the model is and how it performs on the data in both the training and testing sets. As previously said, twenty % goes into testing and eighty % goes into teaching. Machine learning employs a variety of matrices to assess and quantify the outcomes of different algorithms.

Several performance matrices are used to assess the performance of machine learning algorithms. However, in this paper, we rely heavily on categorization. As a result, we use measurements that are appropriate for our challenge. When the number is 1 (malignant) for breast cancer prediction, this is a positive example, suggesting that the person has Breast cancer. When the result is 0 (benign), it is a negative instance, indicating that the patient is cancer-free. As a result, we use confusion measures, which are often utilized in classification problems.

4.1 Confusion Matrix of SVM

We trained the model using SVM algorithms and analyzed the results using a confusion matrix. (Table 4) describe the SVM confusion matrix.

Table 4. Confusion matrix of SVM

	Actual Negative	**Actual Positive**
Predicted Negative	74	0
Predicted Positive	2	38

(Table 5) describe the SVM classification report in which accuracy, precision, recall, F1 score and error rate result are shown.

Table 5. Classification Report of SVM

Accuracy	**Precision**	**Recall**	**F1_Score**	**Specificity**	**Error Rate**
97.45%	0.86	0.89	0.88	0.99	0.0154

4.1.1 Confusion Matrix of K-NN

The k-NN accuracy is controlled using the Confusion matrix and the recall, precision, f1 score, and specificity parameters. (Table 6) explain the K-nearest neighbor of the Confusion matrix report below.

Table 6. Confusion matrix of KNN

	Actual Negative	**Actual Positive**
Predicted Negative	54	2
Predicted Positive	3	55

(Table 7) describe the KNN classification report in which accuracy, precision, recall, F1 score and error rate result are shown.

Table 7. Classification Report of KNN

Accuracy	**Precision**	**Recall**	**F1_Score**	**Specificity**	**Error Rate**
91.44%	0.83	0.84	0.83	0.82	0.0520

4.1.2 Confusion Matrix of Logistic regression

Logistic regression is similar to linear regression; however, the main distinction is that logistic regression is used for categorization based on independent and dependent variables. (Table 8) we assess the logistic regression, which gives the confusion matrix report given below.

Table 8. Confusion Matrix of LR

	Actual Negative	**Actual Positive**
Predicted Negative	70	0
Predicted Positive	4	40

(Table 9) describe the logistic regression classification report in which accuracy, precision, recall, F1 score and error rate result are shown.

Table 9. Classification report of LR

Accuracy	Precision	Recall	F1_Score	Specificity	Error Rate
94.87%	0.88	0.85	0.85	0.8923	0.0443

4.1.3 Confusion Matrix of Random Forest

Random forest is another key algorithm in machine learning or the backbone. Because it can solve classification and regression difficulties. (Table 10) we utilize the random forest's confusion report.

Table 10. Confusion Matrix of RF

	Actual Negative	Actual Positive
Predicted Negative	60	2
Predicted Positive	0	44

(Table 11) describe the random forest classification report in which accuracy, precision, recall, F1 score and error rate result are shown.

Table 11. Classification report of RF

Accuracy	Precision	Recall	F1_Score	Specificity	Error Rate
94.52%	0.89	0.89	0.87	0.8357	0.0454

4.1.4 Confusion Matrix of Artificial Neural Network

The humanoid mind stimulates Ann. Ann performs a similar purpose in the human psyche. Because ANNs include continuous values, such as regression, the confusion matrix cannot accept both binary and continuous parameters, hence we utilize an accuracy graph here. Other parameters stay the same as in earlier algorithms. (Table 12) present a clear image of ANN accuracy and tell us how accurate it is, as well as explain the Confusion matrix of ANN.

Table 12. Confusion Matrix of ANN

	Actual Negative	Actual Positive
Predicted Negative	55	1
Predicted Positive	1	57

(Table 13) describe the ANN classification report in which accuracy, precision, recall, F1 score and error rate result are shown.

Table 13. Classification report of ANN

Accuracy	Precision	Recall	F1_Score	Specificity	Error Rate
97.442%	0.86	0.89	0.87	0.987	0.0149

4.2 Comparative Analysis of ML Algorithms

The goal of this portion of our thesis is to examine the outcomes of all applied models. In this section, we will compare accuracy, precision, recall, f1-Score, specificity, and error rate. The sole accuracy utilized to evaluate a model's performance is not a single parameter on which we concentrated. We want a perfect model that is an excellent predictor, therefore we concentrate on the many parameters that we covered earlier. The clear outcomes of all the fall algorithms that we built in this study are shown in (Table 14).

Table 14. Analysis of All Algorithms

Algorithms	Accuracy	Precision	Recall	F1-Score	Specificity	Error rate
SVM	97.45%	0.86	0.89	0.88	0.89	0.0154
k-NN	91.44%	0.83	0.84	0.83	0.82	0.0520
LR	94.87%	0.88	0.85	0.85	0.8923	0.0443
RF	94.52%	0.89	0.89	0.87	0.8357	0.0454
ANN	97.447%	0.86	0.89	0.87	0.887	0.0149

Figure 3, below explain the accuracy, precision, recall and F1 score of the different algorithms. In this figure comparison of SVM, KNN, LR, RF and ANN and SVM shows better performance as compared to another algorithm then ANN show better performance.

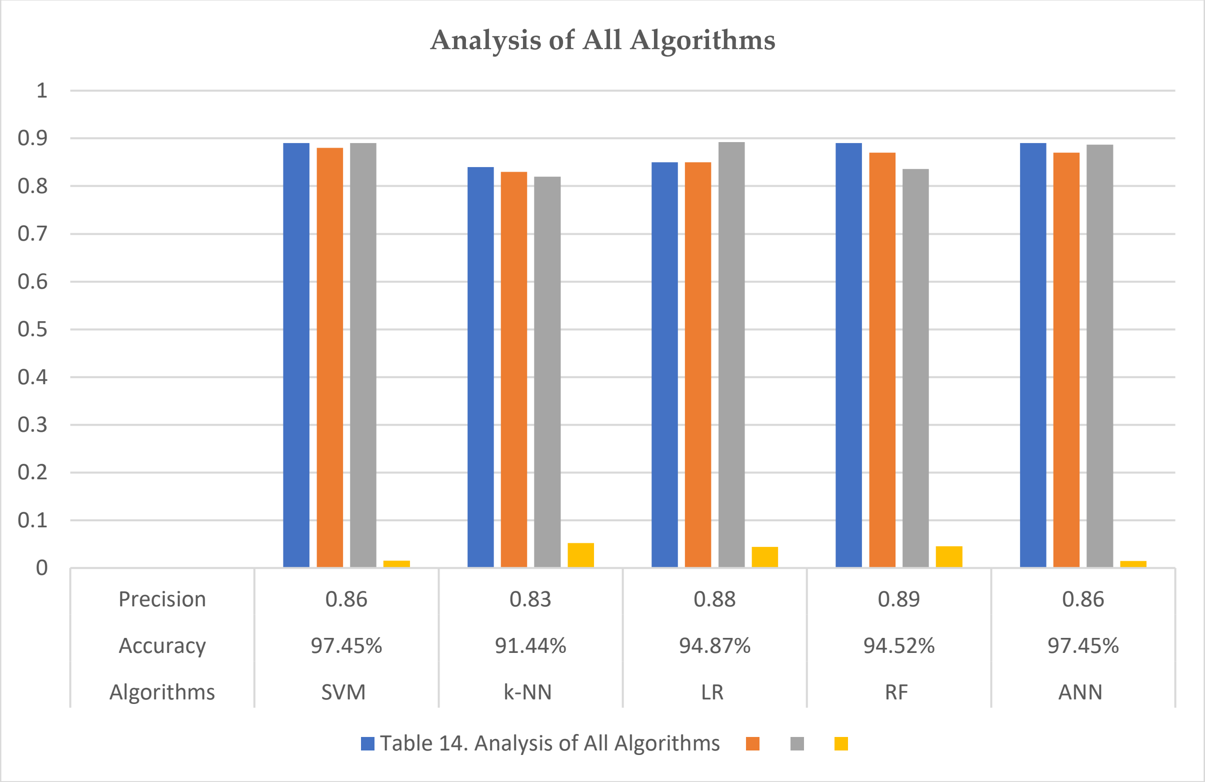


Figure 3. Accuracy and Precision comparison

4.3 Model Performances with fewer Features

The models shown above utilize the whole dataset and all of the thirty characteristics. However, in this part, we assess the correlation between traits and use those that are unique and have the least amount of similarity. Following the discovery of the correlation, the total number of characteristics that will be employed in algorithms will be sixteen rather than thirty. Using a heatmap, we discover a correlation. The first image depicts the relationship between thirty characteristics. The correlation between characteristics or input data is defined by the color bar.

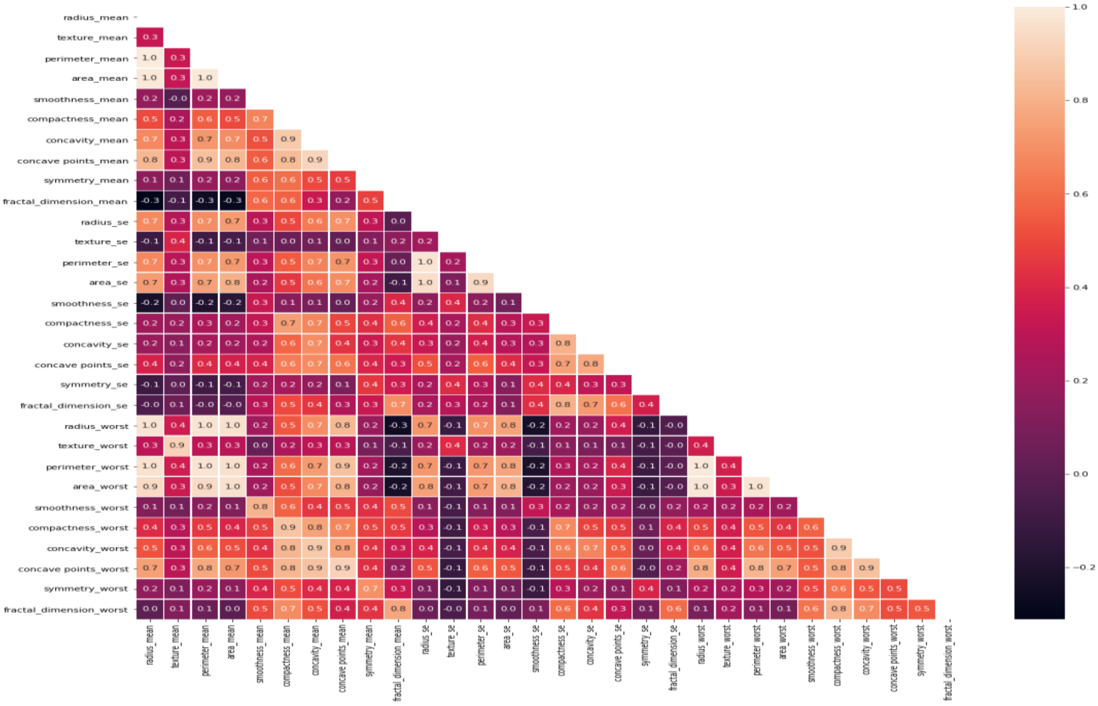


Figure 4. Heat map of 30 features

This (figure 4) describes the heat map the potential features and algorithms that will be used in this figure. 30 characteristics are extracted to work, and the dark region represents the features that are employed; also, the light black area indicates the features that are substituted.

4.3.1 Confusion Matrix of SVM

The confusion matrix is used to determine how accurate our models are. The support vector machine is capable of both regression and classification. (Table 15) The SVM confusion matrix with fewer features is presented below.

Table 15. Confusion Matrix of SVM with Selected Features

	Actual Negative	Actual Positive
Predicted Negative	53	3
Predicted Positive	2	34

(Table 16) The classification report of SVM with 16 features is shown below here.

Table 16. Classification Report of SVM with Selected Features

Accuracy	Precision	Recall	F1_Score	Specificity	Error Rate
96.43%	0.85	0.88	0.86	0.96	0.0314

4.3.2 Confusion matrix of KNN

A confusion matrix, also known as an error matrix, is used in the field of machine learning and precisely the issue of statistical categorization. A confusion matrix is a table that is often used to represent the performance of a classification model (or "classifier") on a set of test data for which the real values are known. (Table 17) The KNN confusion matrix is illustrated below.

Table 17. Confusion Matrix of KNN with Selected Features

	Actual Negative	Actual Positive
Predicted Negative	72	2
Predicted Positive	0	33

(Table 18) classification report of KNN with selected features:

Table 18. Classification Report of KNN with Selected Features

Accuracy	Precision	Recall	F1_Score	Specificity	Error Rate
92.36%	0.85	0.86	0.87	0.88	0.0823

4.3.3 Confusion matrix of Logistic regression

Logistic regression is similar to linear regression, except it is used to solve classification problems. The logistic regression confusion matrix (Table 19) is provided below.

Table 19. Confusion Matrix of LR with Selected Features

	Actual Negative	Actual Positive
Predicted Negative	53	4
Predicted Positive	5	50

(Table 20) explain the Classification report of logistic regression.

Table 20. Classification Report of LR with Selected Features

Accuracy	Precision	Recall	F1_Score	Specificity	Error Rate
95.69%	0.89	0.86	0.86	0.87	0.0842

4.3.4 Confusion matrix of Random Forest

Random forest is based on assembling learning. We make decisions based on a majority vote in this case. Explain the Random Forest Confusion Matrix (Table 21).

Table 21. Confusion Matrix of Random Forest with Selected Features

	Actual Negative	Actual Positive
Predicted Negative	54	4
Predicted Positive	3	53

(Table 22) explain the classification report of Random Forest.

Table 22. Classification Report of RF with selected Features

Accuracy	Precision	Recall	F1_Score	Specificity	Error Rate
95.45%	0.88	0.85	0.86	0.85	0.0858

4.3.5 Confusion Matrix of Artificial Neural Network

Ann draws inspiration from biological brain networks. It can solve both regression and classification problems. Because we have continuous values during testing, displaying the confusion matrix of a neural network is tricky. So (Table 23) explains the ANN Confusion matrix.

Table 23. Explain the confusion Matrix of ANN with Selected Features

	Actual Negative	Actual Positive
Predicted Negative	59	5
Predicted Positive	3	51

(Table 24) explain the classification report of ANN.

Table 24. Classification Report of ANN with selected Features

Accuracy	Precision	Recall	F1_Score	Specificity	Error Rate
96.428%	0.85	0.88	0.86	0.956	0.0312

4.4 Comparative Analysis of ML Algorithms

The purpose of this component of our research is to look at the results of all installed models. We will compare accuracy, precision, recall, f1-Score, Specificity, and error rate in this section. The only accuracy used to assess a model's performance is not a single parameter that we focused on. We want a perfect model that is a great predictor, so we focus on the various parameters we discussed before. Table 25 shows the clear results of all the Algorithms used in this paper.

Table 25. Analysis of All Algorithms with 16 selected Features

Algorithms	Accuracy	Precision	Recall	F1-Score	Specificity	Error rate
SVM	96.43%	0.85	0.88	0.86	0.86	0.0314
k-NN	92.36%	0.85	0.86	0.87	0.88	0.0823
LR	95.69%	0.89	0.86	0.86	0.87	0.0842
RF	95.45%	0.88	0.85	0.86	0.85	0.0858
ANN	96.43%	0.85	0.88	0.86	0.856	0.0312

A graphical Representation of All Algorithms with 16 Selected features is shown in (figure 4). In which comparisons are done on the base of accuracy, precision, F1 score and the error rate of different algorithms.

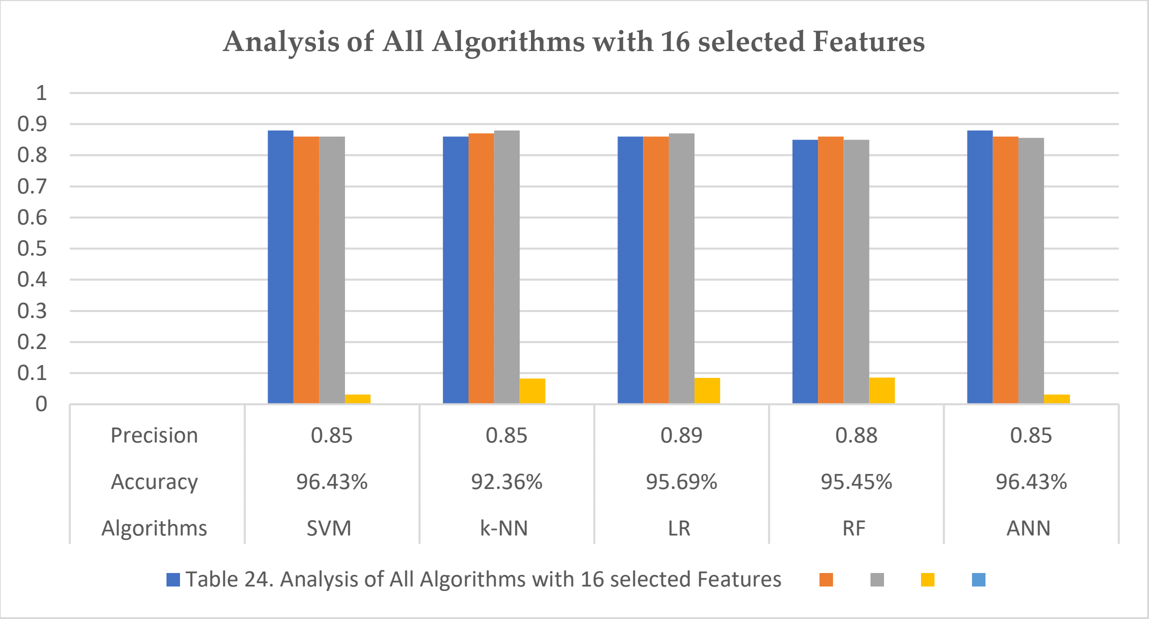


Figure 4. Analysis of All Algorithms with 16 selected Features

The suggested hybrid technique uses the best classifiers for reduced features, such as the Support vector machine (SVM), for accuracy and precision, and produces unaltered accuracy with decreased features, implying that SVM may be employed for greater accuracy with fewer features and less computation time. And KNN displays poorer accuracy after feature removal. For F1 scores, logistic regression (LR) produces the best results when characteristics are decreased. Although the error rate for most classifiers has grown, it remains below acceptable limits for SVM and LR.

Table 26. Analysis of All Algorithms with 9 selected Features

Algorithms	Accuracy	Precision	Recall	F1-Score	Specificity	Error rate
SVM	96.43%	0.86	0.88	0.87	0.86	0.0304
k-NN	91.45%	0.84	0.84	0.86	0.84	0.0823
LR	93.45%	0.82	0.82	0.86	0.83	0.0842
RF	94.34%	0.83	0.82	0.86	0.82	0.0858
ANN	96.42%	0.86	0.88	0.87	0.856	0.0309

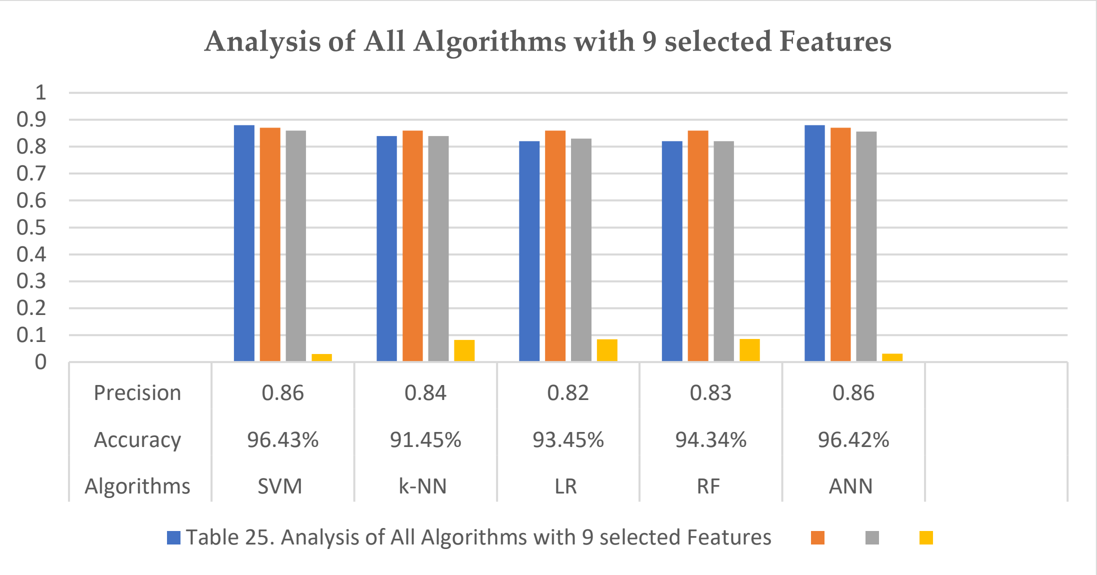


Figure 5. Analysis of All Algorithms with 9 selected Features

A graphical Representation of All Algorithms with 9 Selected features is shown in (figure 5). In which comparison analysis is done on the base of accuracy, precision, F1 score and the error rate of different algorithms.

4.5 Computational Analysis:

(Table 27) is explained that a 50% reduction in time is a tremendous accomplishment, particularly when accuracy and precision are maintained to ensure high dependability with fewer characteristics. With fewer features and better time management, the suggested method has produced strong confidence.

Table 27. Computational Time

Classifiers	9 Features	16 Features	Time saved
SVM	0.95	0.48	0.47
KNN	0.74	0.54	0.2
ANN	0.53	0.33	0.2
LR	0.76	0.28	0.48
RF	0.94	0.49	0.45

5. Conclusion and Future Work

This study compares several machine learning classifiers and how they are implemented. For cancer identification with fewer features, we use support vector machines, K-nearest neighbors, random forests, logistic regression, and artificial neural networks. While maintaining accuracy, precision, F1 score, and recall far beyond the necessary level of diagnosis of breast cancer, a supervised learning technique is used. The best classifiers for decreased features were included in the suggested hybrid technique, including the Support vector machine (SVM), which achieved unaltered accuracy with feature reduction. As a result, SVM may be used for improved accuracy with fewer features and faster computation. Additionally, KNN exhibits decreased accuracy after feature reduction. Reduced feature logistic regression (LR) yields the best results for the F1 score. Even though error rates have grown for the majority of classifiers, SVM and LR continue to operate within acceptable bounds. A 50% reduction in time is a considerable accomplishment, particularly when accuracy and precision are maintained and excellent dependability is provided with fewer features. With fewer features and better time management, the suggested method has produced strong confidence.

Wisconsin Breast Cancer Dataset is utilized for result evolution for breast cancer diagnostics. The outstanding potential of ML techniques to improve classifier accuracy for diagnosis has been shown. Numerous algorithms have used WBCD to attain excellent accuracy. Better algorithms must yet be created, however. Although not the primary assessment element, categorization accuracy is quite important. Particular algorithms have particular mechanisms and consider particular factors. SVM and artificial neural networks both have an accuracy of 97.45% throughout the whole dataset, with an error rate of 0.0154. The third algorithm, Random Forest, came in at 94.53%, which is a respectable result. The k-NN method, with a 91.44% accuracy rate, has the poorest accuracy out of the other four. If we employ PCA, which reduces the parameters used in feature selection, this might be enhanced if we have a big dataset.

Despite our success in collecting results with a respectable degree of accuracy, there have been certain obstacles that have arisen along the course of this research. Although it is undeniable that algorithms should have been thoroughly assessed with huge datasets, the initial issue was the lack of sufficiently large datasets. A large dataset's accessibility may be used to evaluate how quickly PCA-based methods operate. In reality, the research lacks dynamic structure, but the model we used to get better results would incorporate dynamic interaction between functions in more advanced models. Since our dataset is quite dated, more statistical parameters and advanced technologies would have been available to acquire more reliable numerical results. This will also check our research to see if we were able to define the right parameters from our present and potential datasets in the study. To assess and continue our search for the best predictive model, we would want to test more algorithms in contrast to the models we have already employed.

Data Availability: Statement: The authors declare that all data supporting the findings of this study are available within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Abiodun, M.K., et al. Comparing the Performance of Various Supervised Machine Learning Techniques for Early Detection of Breast Cancer. in Hybrid Intelligent Systems. 2022. Cham: Springer International Publishing.
2. Singh, G., et al. Breast Cancer Screening Using Machine Learning Models. in 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM). 2022.
3. Amethiya, Y., et al., Comparative analysis of breast cancer detection using machine learning and biosensors. Intelligent Medicine, 2022. 2(2): p. 69-81.
4. Sengar, P.P., M.J. Gaikwad, and A.S. Nagdive. Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction. in 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). 2020.
5. Thomas, T., N. Pradhan, and V.S. Dhaka. Comparative Analysis to Predict Breast Cancer using Machine Learning Algorithms: A Survey. in 2020 International Conference on Inventive Computation Technologies (ICICT). 2020.
6. Asri, H., et al., Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science, 2016. 83: p. 1064-1069.
7. Sakib, S., et al. Breast Cancer Detection and Classification: A Comparative Analysis Using Machine Learning Algorithms. in Proceedings of Third International Conference on Communication, Computing and Electronics Systems. 2022. Singapore: Springer Singapore.
8. Bacha, S. and O. Taouali, A novel machine learning approach for breast cancer diagnosis. Measurement, 2022. 187: p. 110233.
9. Bayrak, E.A., P. Kırcı, and T. Ensari. Comparison of Machine Learning Methods for Breast Cancer Diagnosis. in 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT). 2019.
10. Rasool, A., et al. Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis. International Journal of Environmental Research and Public Health, 2022. 19, DOI: 10.3390/ijerph19063211.
11. Bazazeh, D. and R. Shubair. Comparative study of machine learning algorithms for breast cancer detection and diagnosis. in 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA). 2016.
12. Ramesh, S., et al., Segmentation and classification of breast cancer using novel deep learning architecture. Neural Computing and Applications, 2022. 34(19): p. 16533-16545.
13. Benbrahim, H., H. Hachimi, and A. Amine. Comparative Study of Machine Learning Algorithms Using the Breast Cancer Dataset. in Advanced Intelligent Systems for Sustainable Development (AI2SD'2019). 2020. Cham: Springer International Publishing.
14. Binsaif, N., Application of Machine Learning Models to the Detection of Breast Cancer. Mobile Information Systems, 2022. 2022: p. 7340689.
15. Ogundokun, R.O., et al. Medical Internet-of-Things Based Breast Cancer Diagnosis Using Hyperparameter-Optimized Neural Networks. Future Internet, 2022. 14, DOI: 10.3390/fi14050153.
16. Dinesh, P. and K. P, Medical Image Prediction for Diagnosis of Breast Cancer Disease Comparing the Machine Learning Algorithms: SVM, KNN, Logistic Regression, Random Forest, and Decision Tree to Measure Accuracy. ECS Transactions, 2022. 107(1): p. 12681.
17. Egwom, O.J., et al. An LDA&ndash;SVM Machine Learning Model for Breast Cancer Classification. BioMedInformatics, 2022. 2, 345-358 DOI: 10.3390/biomedinformatics2030022.
18. Nomani, A., et al., PSOWNNs-CNN: A Computational Radiology for Breast Cancer Diagnosis Improvement Based on Image Processing Using Machine Learning Methods. Computational Intelligence and Neuroscience, 2022. 2022: p. 5667264.
19. Gayathri, B.M. and C.P. Sumathi. Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer. in 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). 2016.
20. Houssein, E.H., M.M. Emam, and A.A. Ali, An optimized deep learning architecture for breast cancer diagnosis based on improved marine predators algorithm. Neural Computing and Applications, 2022. 34(20): p. 18015-18033.
21. Jasti, V.D.P., et al., Computational Technique Based on Machine Learning and Image Processing for Medical Image Analysis of Breast Cancer Diagnosis. Security and Communication Networks, 2022. 2022: p. 1918379.
22. Jabeen, K., et al. Breast Cancer Classification from Ultrasound Images Using Probability-Based Optimal Deep Learning Feature Fusion. Sensors, 2022. 22, DOI: 10.3390/s22030807.
23. Nanglia, S., et al., An enhanced Predictive heterogeneous ensemble model for breast cancer prediction. Biomedical Signal Processing and Control, 2022. 72: p. 103279.
24. Madani, M., M.M. Behzadi, and S. Nabavi The Role of Deep Learning in Advancing Breast Cancer Detection Using Different Imaging Modalities: A Systematic Review. Cancers, 2022. 14, DOI: 10.3390/cancers14215334.
25. Monirujjaman Khan, M., et al., Machine Learning Based Comparative Analysis for Breast Cancer Prediction. Journal of Healthcare Engineering, 2022. 2022: p. 4365855.
26. Mangasarian, O.L., W.N. Street, and W.H. Wolberg, Breast Cancer Diagnosis and Prognosis Via Linear Programming. Operations Research, 1995. 43(4): p. 570-577.
27. Michael, E., et al., An Optimized Framework for Breast Cancer Classification Using Machine Learning. BioMed Research International, 2022. 2022: p. 8482022.