*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 7s*
*DOI: https://doi.org/10.17762/ijritcc.v11i7s.7527*
***Article Received**: 24 April 2023 **Revised:** 14 June 2023 **Accepted:** 28 June 2023*

# Breast Cancer Detection by Extracting and Selecting Features Using Machine Learning

**[1,*] Priyanka M. Tambat, [2]Dr. Sohel A. Bhura, [3]Dr. Salim Y. Amdani, [4]Dr. Suresh S. Asole**
[1,3,4] Babasaheb Naik College of Engineering, Pusad, Maharashtra, India
[2]Jhulelal Institute of Technology, Nagpur, Maharashtra, India
[1]mail.priyankaaher@gmail.com, [2]sabhura@rediffmail.com, [3]drsalimamdani@gmail.com, [4]suresh_asole@yahoo.com

**Abstract:** The cancer of the breast is a significant cause of female death worldwide, but especially in developing countries. For better results and higher survival rates, early diagnosis and screening are crucial. Machine learning (ML) methods can aid in the initialdiscovery and diagnosis of breast cancer by choosing the most informative elements from medical data and eliminating irrelevant ones. The approach of feature extraction involves taking unstructured data and extracting a representative set of characteristics that may be used to classify or forecast data. The aim is to decrease the dimensionality of the feature space while upholding or even refining the accuracy of the ML model. An artificial intelligence model is developed on the given features to categorize mammography images into benign and malignant groups. Different supervised learning techniques, including support vector machines, random forests, and artificial neural networks, are employed and contrasted in order to select the best-performing model. This research offers a comprehensive framework for utilizing machine learning methods to detect breast cancer. The technique demonstrates how it might assist radiologists in the early detection of breast cancer by effectively extracting and selecting critical characteristics that could improve patient outcomes and potentially save lives.

**Keywords:** Feature Selection, Feature Extraction method, Support Vector Machine, Random forest, Recursive Feature Elimination.

## I. INTRODUCTION

By extracting and choosing essential information from the patient's data, automatic learning algorithms can help doctors identify and diagnose breast cancer. Extraction and selection of features are critical steps in detecting relevant patterns and qualities in medical data[5]. Machine learning algorithms can learn to identify between malignant and non-cancerous instances by extracting useful information from multiple sources such as mammograms, clinical records, and genetic data. Feature extraction techniques convert raw data into a compact and representative set of characteristics, making classification or prediction jobs more efficient and accurate [4].

By finding the most useful and discriminative features, feature selection approaches handle the difficulty of high-dimensional data [1]. Breast cancer arises when the unregulated development of cells in the breast's fatty and fibrous tissues results in the production of malignant tumors. These malignant cells have the ability to spread beyond the primary tumor, moving through increasing severity stages. Breast cancer's aggressive nature contributes to its rank as one of the most fatal diseases in contemporary times [2].

Breast abnormalities can be found using a number of methods, including imaging, medical exams, and self-examination. By extracting key information from various medical datasets, machine learning algorithms can more effectively and precisely diagnose breast cancer patients [3].However, mammography has its limitations in specific situations, especially for people with dense pectoral tissue. Additionally, there is a large rise in the danger of ionizing radiation exposure, which is important for young women. Furthermore, it can be challenging to find tumors with a diameter of less than 2 mm utilizing mammography. These drawbacks highlight the value of ongoing research to enhance mammography imaging for the early diagnosis of breast cancer [10].

In direction to fully utilize the potential of ML models, this work provides a novel method for breast cancer showing that places a strong emphasis on achieving high accuracy. The recommended approach makes the following crucial contributions:

- The method use feature selection techniques to identify the most informative and relevant characteristics from a variety of datasets. By picking the most discriminative traits.
- The goal of this study is to determine how intricate features and hand-crafted traits affect breast cancer prediction. The study focuses on extracting complex characteristics with CNN
- A machine learning system that is effective in predicting colon cancer is proposed. ML classifiers are used by the model to increase prediction

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 7s*
*DOI: https://doi.org/10.17762/ijritcc.v11i7s.7527*
***Article Received**: 24 April 2023 **Revised:** 14 June 2023 **Accepted:** 28 June 2023*

accuracy. The combined judgment of these classifiers is employed in a voting procedure to produce the final forecast.

The paper has been organized as: Section II gives a succinct overview of the existing breast cancer detection literature, noting research limitations in the field. Section III describes the dataset used in this study, providing useful insights for future research. Section IV describes the suggested methodology in depth, including the feature extraction and selection methods used. Section V presents the study's findings, followed by Section VI conclusion to the research.

## II. REVIEW OF LITERATURE

Numerous studies on breast cancer detection have been undertaken, and computer-aided diagnostics (CAD) has emerged as a critical tool in the early phases of diagnosis.Health analytics, however, has challenges when analysing large and varied healthcare data. The development of CAD and AI has paved the way for more precise and accurate systems in medical applications, especially when working with sensitive medical data.

Even in developed nations, stomach cancer remains the leading cause of death. A significant portion of CAD and decision support research is focused on tumor detection, particularly with regard to this disease. Automated learning techniques may help to detect gastrointestinal cancer. While some research have used single strategies to get accurate findings, there is a rising interest in using ensemble models for greater performance. This section of the paper gives a detailed evaluation of the most recent and cutting-edge breast cancer screening tools that use machine learning methodologies.

The effectiveness of the k-nearest-neighbours (KNN) and Naive Bayes (NB) methods for determining the type of breast cancer was investigated by Amrane et al. [12]. The researchers utilized K-fold cross-validation to assess the models' performance after classifying the tumors as benign or malignant. The testing results show that KNN achieved a fantastic precision of 97.51% for the binary categorization challenge.

Nawaz et al. [14] concentrated on multiclass classification of breast cancer by dividing tumors into three subclasses. To do this, the authors used a Convolutional Neural Network (CNN) on histopathology pictures from the BreakHis dataset. They classified data using a deep CNN model, and their 95.4% accuracy rate was astounding. Singh et al. [15] used auto-encoders and other machine learning algorithms to predict breast cancer. The study also included an unsupervised auto-encoder model developed exclusively for breast cancer detection. The authors concentrated on creating a concise feature representation of breast cancer. Notably, the auto-encoder model beat the other classifiers employed in the study, scoring 98.4% precision and recall.

The fuzzy logic system's rules were built on top of the optimal membership functions that were discovered in the [16] study using a subset of the data. The performance of cancer diagnosis was enhanced by combining these two technologies, algorithmic genetics and a fuzzy logic system. These studies show that alternate ways to breast cancer detection are effective. Singh et al. demonstrated the use of auto-encoders by getting good precision and recall scores. Allison Murphy's study, on the other hand, used the GFS-TSK technique to improve cancer detection performance by utilizing genetic algorithms and fuzzy logic. These revolutionary procedures show the potential of creative methodology in improving breast cancer diagnosis, presenting promising routes for future research and contributing to medical diagnostic improvement.

Table 1: Comparative Summary of existing method

| Paper | Algorithm | Types of Dataset | Remark |
|---|---|---|---|
| [14] | Naïve Bays and K Nearest neighbor | Cancer Dataset | KNN obtained an accuracy of 97.5%. |
| [15] | The SVM, DT, and K Nearest Neighbor | Cancer of the breast | SVM performed well, achieving 98.1% accuracy. |
| [16] | Convolution neural Network | BreakHis dataset | CNN achieved an accuracy of 95.4%. |
| [13] | Scheme Autoencoder | Cancer of the breast Wisconsin dataset | Other models were outperformed by the proposed method. |
| [18] | Fuzzy Logic Systems | Cancer of the breast Wisconsin dataset | When paired with fuzzy logic, the genetic algorithm outperforms. |
| [17] | XGBoost Method | Cancer of the breast Wisconsin dataset | Using 13 characteristics, the proposed method achieved 97.7% accuracy. |
| [18] | GBM, XGBoost, LightGBM | Cancer of the breast Wisconsin dataset | LightGBM produced reliable findings in the categorization of breast tumors. |
| [19] | Logistic Rregression, Decision Tree, K Nearest neighbor and Naïve Bays | Cancer of the breast Wisconsin dataset | LR received the best results. |
| [20] | ML algorithms | Mammogram | Hybrid models produced reliable findings. |

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 7s*
*DOI: https://doi.org/10.17762/ijritcc.v11i7s.7527*
***Article Received**: 24 April 2023 **Revised:** 14 June 2023 **Accepted:** 28 June 2023*

| | | | |
|---|---|---|---|
| [21] | K Nearest neighbor (KNN) , DT and (SVM) Support Vector machine | Breast thermal images | To improve image quality, the Firefly algorithm was used. |
| [22] | SVM-coarse Gaussian and SVM-cubic | thermal pictures of the breast | The accuracy of the suggested method was 93.5%. |
| [23] | AOL, RetinaNet | Mammogram | The study's breadth was constrained because it only tested five images. |
| [24] | Convolution neural Network | Cancer of the breast Wisconsin dataset | Individual models were outperformed by the proposed layered ensemble technique. |

## III. BREAST CANCER DATASET DESCRIPTION

The documentation of cancer at UCI lists 569 cases and thirty characteristics. The smoothness, radius, and texture of the samples were used to distinguish between benign and malignant tumors during sample collection. According to these standards, tumors will be categorized as benign or malignant. With the use of this data base, it is possible to create and test an automatic learning model for classifying and identifying breast cancer.

Table 2: Additional feature summary for Breast Cancer Wisconsin (BCW) dataset [25]

| Feature | Description |
|---|---|
| ID number | Unique identifier for each cell nucleus |
| Diagnosis | Malignant (M) or Benign (B) |
| Radius | Average distance between the perimeter's points and the center |
| Texture | Standard deviation of values in grayscale |
| Perimeter | around the cell nucleus |
| Area | space around a cell's nucleus |
| Smoothness | Radius length variations on a local level |
| Compactness | Calculated as (perimeter^2 / area) - 1.0 |
| Concavity | severity of the contour's concave areas |
| Concave points | The proportion of the contour's concave areas. |
| Symmetry | Nucleus of the cell's symmetry |
| Fractal dimension | "Coastline approximation" - 1, representing the complexity of the cell nucleus boundary |

These elements indicate different aspects of the cell nuclei shown in the photograph. They provide important information on the features and characteristics of the nuclei, which is required for the analysis and classification of breast masses. The dataset contains critical features that help academics and practitioners comprehend the cellular composition and structure, allowing them to generate insights and models for breast cancer diagnosis and categorization.

Table 3: Description of Dataset

| Sr. No | Characteristic | Associated for | Area | Attribute Type | Attributes | Records |
|---|---|---|---|---|---|---|
| 1 | Multivariate | Classification | Life | Real | 30 | 569 |

## IV. PROPOSED SYSTEM

The first stage of the process involves gathering data by extracting microscopic characteristics related with the breast from cell nuclei. Following retrieval, these traits are preprocessed, which entails converting categorical features to numeric form using a label encoder. It's important to note that the dataset doesn't have any null values. After preprocessing, the processed microscopic features are created. This division is performed using the train-test validation method from the Sklearn package. The training set is used to obtain deep convoluted features. These characteristics are obtained by employing machine learning method. Deep learning algorithms enable the extraction of extremely useful and discriminative features from training data.Following this data collecting and preprocessing approach yields valuable microscopic breast features that can be used for later analysis and classification activities.

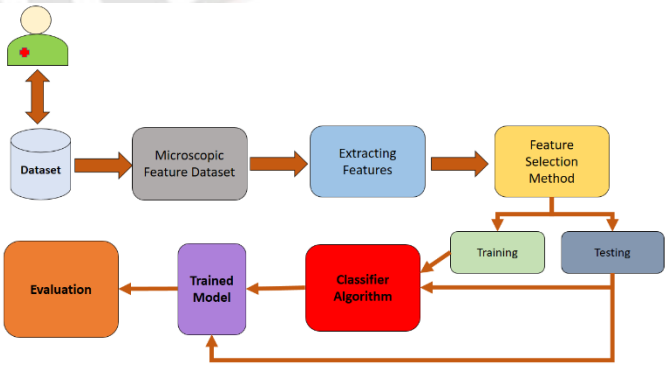


Figure 1: Proposed system architecture of Feature selection and Classification

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 7s*
*DOI: https://doi.org/10.17762/ijritcc.v11i7s.7527*
***Article Received**: 24 April 2023 **Revised:** 14 June 2023 **Accepted:** 28 June 2023*

**1. RFE algorithm:**

The Recursive Feature Elimination (RFE) algorithm is a attribute selection strategy extensively used in machine learning for feature extraction. It is an iterative approach that seeks to identify the most essential aspects in a dataset by removing less relevant features recursively.

The RFE algorithm begins by training a model on all of the features and rating them in order of importance. The method then discards the least important characteristics and restarts the model with the smallest possible set of characteristics. This tactic is repeatedly used until the necessary number of characteristics or the level of relevance of characteristics is reached.

Algorithm:

- Initialization: Random positions and velocities are assigned to each particle in the swarm.
- Evaluation: The objective function is evaluated for each particle, and the fitness value is calculated based on the function's output.
- Finding pbest: If the current fitness value is better than the personal best (pbest) value for a particle, the fitness value is updated as the new pbest.

$$f(x) = \begin{cases} x, & if\ x > 0 \\ ax, & Otherwise \end{cases}$$

- Finding gbest: If the pbest value of any particle is better than the global best (gbest) value, the gbest value is updated.
- Updating the position and velocity: A unique equation is used to update the velocity of each particle. The particle's location is then updated using the new velocity.
- Terminating Criteria: The process starts when a final condition is met. It is possible that this illness will progress to a certain number of repetitions or a physical limit. The algorithm returns to phase 2 and repeats the process if the termination condition is not met.

**2. (FSS) Feature subset selection:**

The process of choosing a crucial subset of characteristics from the initial set of characteristics is known as the selection of sub-conjunto of characteristics. This set of characteristics has the potential to accurately reflect the data and boost the effectiveness of automatic learning models.

**Step1:** $\boldsymbol{Calculate\ the\ mutual\ information\ as\ score\ for\ between\ all\ features}$ $(\boldsymbol{Fi\ \in f})$ $\boldsymbol{and\ the\ target\ class}$ $(\boldsymbol{c})$

$\boldsymbol{Step\ 2}$: $\boldsymbol{Select\ the\ feature\ with\ the\ largest\ score}$ $(\boldsymbol{e.g.\ argmax})$ $\boldsymbol{add\ it\ to\ the\ set\ of\ selected\ features}$ $(\boldsymbol{S})$

$\boldsymbol{Step\ 3}$: $\boldsymbol{Calculate\ the\ score\ which\ might\ be\ derived\ from\ the\ mutual\ information}$

$\boldsymbol{Step\ 4}$: $\boldsymbol{Select\ the\ feature\ with\ the\ largest\ score\ and\ add\ it\ to\ the\ set\ of\ select\ features}$

$\boldsymbol{Step\ 5}$: $\boldsymbol{Repeat\ 3.\ and\ 4.\ until\ a\ certain\ number\ of\ features\ is\ selecte}$

**3. Support Vector Machine (SVM):**

The supervised learning subcategory of machine learning incl udes the sophisticated and popular Support Vector Machine (S VM) method. It is typically used for problems involving classi fication and regression.Finding the ideal hyperplane that maxi mally separates the data points belonging to various classes is the basic tenet of SVM. The data points closest to the decision boundary, called support vectors, are used to calculate the hyp erplane. SVM tries to achieve the largest margin between the s upport vectors of various classes, enhancing generalization an d making it resistant against noise. This model is trained to loc ate the best hyper plane that maximally separates the feature v ectors according to their class labels.

**Testing Phase (b)**

i. Test Data: Fresh feature vectors that have been flattened fro m unused photos serve as the test data.

ii. Classification: The test data are divided into the appropriate classes using the trained SVM model.

**4. Random Forest (RF):**

The term "random forest" denotes that the algorithm creates a structure like a forest made up of numerous individual trees. Given that it integrates the results of several algorithms, it is referred to as an ensemble algorithm.A vector of entry (x) containing the values of various characteristics assessed for each instance of formation is used to construct a number of RF regression trees. Results are analyzed and quantified. The RF regression predictor can be stated as follows after creating K trees, indicated as $T(x)K_1$.

$$\int_{rf}^{K} f(x) = 1/K \sum_{k=1}^{K} T(x)$$

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 7s*
*DOI: https://doi.org/10.17762/ijritcc.v11i7s.7527*
***Article Received**: 24 April 2023 **Revised:** 14 June 2023 **Accepted:** 28 June 2023*

Here, RF(x) denotes the anticipated value for the input vector x based on the RF regression model. As indicated by the notation T(x), each distinct regression tree offers a forecast for the input vector x. The final forecast of the RF model is created by averaging (adding up and dividing by K) the predictions from each of the K trees.

Each tree in a random forest utilizes a random selection of features and is trained on a distinct subset of the training data. With less overfitting, the model performs better overall and is more robust thanks to this randomness. To arrive at a final categorization determination during prediction, the random forest integrates the forecasts of each individual tree.

**Algorithm Random Forest:**

***Let D be a training set*** $D = \{(x1, y1), \ldots, (xn, y,)\}$
***Let*** $h = h1(x), h2(x), \ldots,$ ***for an ensemble of weak classifier***
***If each*** $h_k$ ***is a decision tree, the parameters of the tree are defined as*** $\theta = (\theta_{k1}, \theta_{k2,} \ldots\ldots, \theta_{kp,})$
***Each decision tree k leads to a classifier*** $h_k(x) = h(X|\theta_k)$
***Final Classification*** $f(x) =$ ***Majority of*** $h_k(x)$

The random forest approach can handle complicated datasets and capture subtle correlations between attributes by utilizing the diversity of numerous decision trees. It is renowned for its capacity to manage high-dimensional data, deal with missing values, and offer perceptions into the significance of features. Overall, the ensemble nature of the random forest method and its capacity to deliver accurate and dependable results make it a strong and well-liked option for classification problems.

**Performance Indices:**

The accuracy (ACC) is calculated as the percentage of correctly classified instances, whether they are normal or attacks, and is determined by the following formula:

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

The formula for calculating precision (P), which is the proportion of pertinent instances among the identified instances:

$$P = \frac{TP}{(TP + FP)}$$

Recall (R) is calculated as the ratio of the number of relevant instances over the total number of relevant instances discovered:

$$\mathrm{R} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}$$

A statistic called the F1-Score combines recall and precision into a single value. The following formula can be used to determine it as the weighted mean of recall and precision:

$$\mathrm{F1Score} = \frac{(2 * \mathrm{P} * \mathrm{R})}{(\mathrm{P} + \mathrm{R})}$$

In particular, when α = 1, the formula for the F1-Score simplifies. Overall, these formulas allow us to calculate accuracy, precision, recall, and the F1-Score, which are commonly used metrics for evaluating classification performance.

## V. RESULT AND DISCUSSION

Selecting the machine learning characteristics that are most useful for a particular set of data sometimes involves using the recursive characteristic elimination (RFE) technique. Functions by removing the less significant characteristics from the combination until the desired number of characteristics is reached. A model of automatic learning determines the value of each characteristic.The outcome of applying RFE on the dataset is determined by the method employed for feature ranking and elimination. Assume, however, that a random forest technique was used for this purpose.

Table 4: Result of Feature extraction selection Algorithm

| Method | Original Feature | Feature Selection |
|---|---|---|
| REE | 30 | 21 |
| FSS | 30 | 17 |

The proposed machine learning technique evaluated and compared the models using several performance measures. A typical metric called accuracy determines the proportion of accurate forecasts to all forecasts provided. The amount of precise information returned by an ML model is measured by the precision metric, which is utilized in document retrieval. Sensitivity is a measure of how many positive results the machine learning model returned.

Table 5: Accuracy percentage of Dataset for Training and testing using 21 feature

| Method | With Selection of 21 features | |
|---|---|---|
| | Accuracy (Training Set %) | Accuracy (Testing Set %) |
| (SVM) Support Vector Machine | 94.46% | 98.24% |
| (RF) Random Forest | 92.25% | 97.51% |
| (LR) Logistic Regression | 92.55% | 96.80% |
| (DT) Decision Tree | 91.18% | 97.81% |

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 7s*
*DOI: https://doi.org/10.17762/ijritcc.v11i7s.7527*
***Article Received**: 24 April 2023 **Revised:** 14 June 2023 **Accepted:** 28 June 2023*

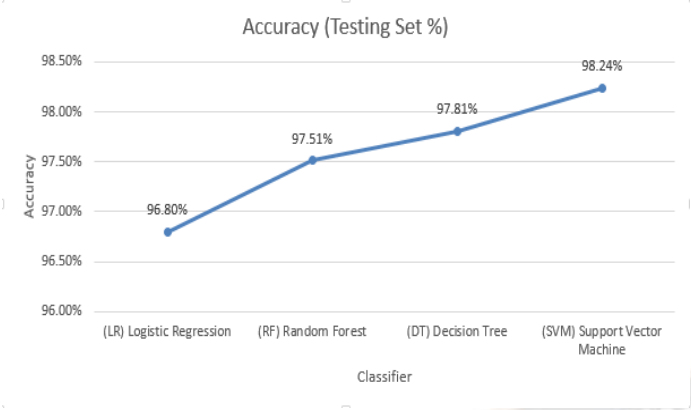


Figure 2: Accuracy of Classifier on Breast Cancer dataset

Recursive Feature Elimination (RFE) is a machine learning feature extraction approach that is often used to choose the most relevant features for a given dataset as discussed in table 5. It operates by removing less important characteristics from the dataset recursively until the desired number of features is obtained. A machine learning model determines the value of each feature.The outcome of applying RFE on the dataset is determined by the method employed for feature ranking and elimination. Assume, however, that a random forest technique was used for this purpose.

Similarly, the RF algorithm worked well, with a training set accuracy of 92.25% and a testing set accuracy of 97.51%. The ability of RF to handle complicated datasets shows its usefulness in capturing underlying patterns and producing accurate predictions.On the training set, the LR algorithm achieved an accuracy of 92.55% and a little lower accuracy of 96.80% on the testing set. When the relationship between the characteristics and the target variable is approximately linear, LR gives good interpretability and performs well.Finally, the DT algorithm obtained a training set accuracy of 91.18% and a testing set accuracy of 97.81%. Decision trees are well-known for their ease of use.

Table 5: Accuracy percentage of Dataset for Training and testing using 17 feature

| Method | Accuracy (Training Set %) | Accuracy (Testing Set %) |
|---|---|---|
| (DT) Decision Tree | 96.18% | 96.81% |
| (LR) Logistic Regression | 94.55% | 97.80% |
| (RF) Random Forest | 96.25% | 98.51% |
| (SVM) Support Vector Machine | 98.46% | 99.24% |

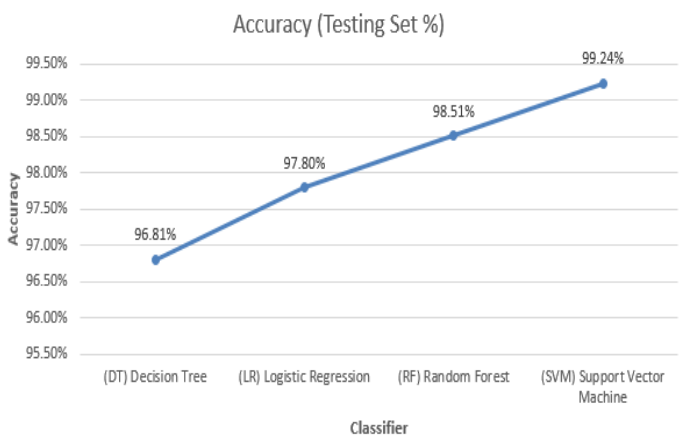


Figure 3:Accuracy percentage of Dataset for testing using 17 feature

Table 5 and figure 3 shows the result of breast cancer dataset, we tested the performance of four machine learning techniques. The Decision Tree (DT) algorithm was 96.18% accurate on the training set and 96.81% accurate on the testing set. Logistic Regression (LR) achieved an accuracy of 94.55% on the training set and 97.80% on the testing set. Random Forest (RF) did even better, with a training set accuracy of 96.25% and a testing set accuracy of 98.51%. With 98.46% accuracy on the training set and 99.24% accuracy on the testing set, the Support Vector Machine (SVM) method surpassed others.

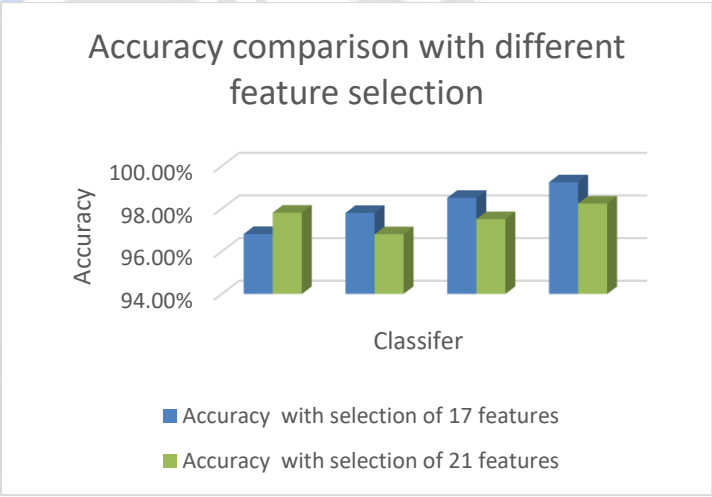


Figure 4: Accuracy comparison with different feature selection

In our experiment, we performed feature selection and assessed how well Decision Tree (DT), Logistic Regression (LR) performed, Random Forest (RF), and Support Vector Machine (SVM). In comparison to a collection of 21 features, we looked at the accuracy achieved with a set of 17 features.LR attained an accuracy of 96.81% with 17 selected features, demonstrating good predictive power. RF achieved 97.80% accuracy, indicating its capacity to handle complicated information successfully. DT scored the

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 7s*
*DOI: https://doi.org/10.17762/ijritcc.v11i7s.7527*
***Article Received**: 24 April 2023 **Revised:** 14 June 2023 **Accepted:** 28 June 2023*

maximum accuracy of 98.51%, demonstrating its ability to grasp complex relationships within data. SVM obtained an astounding 99.24% accuracy, demonstrating its high classification capabilities.LR improved somewhat to 97.81% accuracy when using 21 selected features. RF had a slightly lower accuracy of 96.80%, implying that the extra features did not significantly improve its performance. DT achieved 97.51% accuracy while maintaining a high level of precision. With the increased feature set, SVM achieved a little lower accuracy of 98.24%.Overall, the chosen features had a significant impact on the algorithms' performance, with each algorithm reacting differently to feature selection. According to the results, SVM consistently outperformed DT in terms of robustness to feature selection fluctuations. LR and RF were more sensitive to changes in feature selection but still maintained good accuracy levels in both cases.

## VI. CONCLUSION

In this study, we extracted and selected features from a dataset to employ machine learning methods to detect breast cancer. Proposed method help to improve the predicted accuracy of the models by extracting and choosing key features. We discovered that feature selection had a significant effect on the algorithm's performance. Our results showed that SVM functioned consistently well, with accuracy rates of 99.24% and 98.24% with 17 and 21 selected features, respectively. DT performed well as well, with accuracy ratings of 98.51% and 97.51% for the respective feature sets. LR displayed consistent accuracy rates of 96.81% and 97.81%, while RF demonstrated higher accuracy rates of 97.80% and 96.80%.The findings highlight the utility of machine learning in the identification of breast cancer. We increased the predictive capabilities of the models and achieved high accuracy rates by using feature extraction and selection approaches. The choice of algorithm was also important, as SVM and DT regularly beat the others in this situation.These findings have important significance for the medical industry, as early detection and treatment of breast cancer can be aided by accurate and efficient detection. Machine learning approaches can help medical personnel make more educated decisions, which could lead to better patient outcomes.

## REFERENCES

[1] Abdullah-Al Nahid, Aaron Mikaelian and Yinan Kong “Histopathological breast-image classification with restricted Boltzmann machine along with backpropagation.” (2018).

[2] Syed Jamal Safdar Gardezi, Ahmed Elazab, Baiying Lei and Tianfu Wang "Breast Cancer Detection and Diagnosis Using Mammographic Data: Systematic Review" (2019).

[3] Saleem Z. Ramadan “Methods Used in Computer- Aided Diagnosis for Breast Cancer Detection Using Mammograms: A Review” (2020).

[4] Gradilone, A.; Naso, G.; Raimondi, C.; Cortesi, E.; Gandini, O.; Vincenzi, B.; Saltarelli, R.; Chiapparino, E.; Spremberg, F.; Cristofanilli, M.; et al., “Circulating tumor cells (CTCs) in metastatic breast cancer (MBC)”, Prognosis, drug resistance and phenotypic characterization. Ann. Oncol. 2011, 22, 86–92.

[5] Sun, Y.S.; Zhao, Z.; Yang, Z.N.; Xu, F.; Lu, H.J.; Zhu, Z.Y.; Shi, W.; Jiang, J.; Yao, P.P.; Zhu, H.P. Risk factors and preventions of breast cancer. Int. J. Biol. Sci. 2017, 13, 1387.

[6] Ahmad, F.K.; Yusoff, N. Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier. In Proceedings of the 2013 13th International Conference on Intellient Systems Design and Applications, Salangor, Malaysia, 8–10 December 2013; pp. 121–125.

[7] Robertson, F.M.; Bondy, M.; Yang, W.; Yamauchi, H.; Wiggins, S.; Kamrudin, S.; Krishnamurthy, S.; Le-Petross, H.; Bidaut, L.; Player, A.N.; et al. Inflammatory breast cancer: The disease, the biology, the treatment. CA A Cancer J. Clin. 2010, 60, 351–375.

[8] Hou, R.; Mazurowski, M.A.; Grimm, L.J.; Marks, J.R.; King, L.M.; Maley, C.C.; Hwang, E.S.S.; Lo, J.Y. Prediction of upstaged ductal carcinoma in situ using forced labeling and domain adaptation. IEEE Trans. Biomed. Eng. 2019, 67, 1565–1572.

[9] Loberg, M.; Lousdal, M.L.; Bretthauer, M.; Kalager, M. Benefits and harms of mammography screening. Breast Cancer Res. 2015, 17, 63.

[10] Ishaq, A.; Sadiq, S.; Umer, M.; Ullah, S.; Mirjalili, S.; Rupapara, V.; Nappi, M. Improving the prediction of heart failure patients’ survival using SMOTE and effective data mining techniques. IEEE Access 2021, 9, 39707–39716.

[11] Amrane, M.; Oukid, S.; Gagaoua, I.; Ensari, T. Breast cancer classification using machine learning. In Proceedings of the 2018 Electric Electronics, Computer Science, Biomedical Engineerings’ Meeting (EBBT), Istanbul, Turkey, 18–19 April 2018; pp. 1–4.

[12] Merwe, M. van der, Petrova, M., Jovanović, A., Santos, M., & Rodríguez, M. Text Summarization using Transformer-based Models. Kuwait Journal of Machine Learning, 1(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/141

[13] Obaid, O.I.; Mohammed, M.A.; Ghani, M.K.A.; Mostafa, A.; Taha, F. Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. Int. J. Eng. Technol. 2018, 7, 160–166.

[14] Nawaz, M.; Sewissy, A.A.; Soliman, T.H.A. Multi-class breast cancer classification using deep learning convolutional neural network. Int. J. Adv. Comput. Sci. Appl. 2018, 9, 316–332.

[15] Singh, S.J.; Rajaraman, R.; Verlekar, T.T. Breast Cancer Prediction Using Auto-Encoders. In International Conference on Data Management, Analytics & Innovation; Springer: Singapore, 2023; pp. 121–132.

[16] Murphy, A. Breast Cancer Wisconsin (Diagnostic) Data Analysis Using GFS-TSK. In North American Fuzzy Information Processing Society Annual Conference; Springer: Cham, Switzerland, 2021; pp. 302–308.

[17] Ghosh, P. Breast Cancer Wisconsin (Diagnostic) Prediction. Available online:

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 7s*
*DOI: https://doi.org/10.17762/ijritcc.v11i7s.7527*
***Article Received**: 24 April 2023 **Revised:** 14 June 2023 **Accepted:** 28 June 2023*

https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin +(diagnostic)

[18] Akbulut, S.; Cicek, I.B.; Colak, C. Classification of Breast Cancer on the Strength of Potential Risk Factors with Boosting Models: A Public Health Informatics Application. Med Bull. Haseki/Haseki Tip Bul. 2022, 60, 196–203.

[19] Ak, M.F. A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications. Healthcare 2020, 8, 111.

[20] Kashif, M.; Malik, K.R.; Jabbar, S.; Chaudhry, J. Application of machine learning and image processing for detection of breast cancer. In Innovation in Health Informatics; Elsevier: Hoboken, NJ, USA, 2020; pp. 145–162.

[21] Pattnaik, M. ., Sunil Kumar, M. ., Selvakanmani, S. ., Kudale, K. M. ., M., K. ., & Girimurugan, B. . (2023). Nature-Inspired Optimisation-Based Regression Based Regression to Study the Scope of Professional Growth in Small and Medium Enterprises. International Journal of Intelligent Systems and Applications in Engineering, 11(4s), 100–108. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2576

[22] Dey, N.; Rajinikanth, V.; Hassanien, A.E. An examination system to classify the breast thermal images into early/acute DCIS class. In Proceedings of the International Conference on Data Science and Applications; Springer: Singapore, 2021; pp. 209–220.

[23] Rajinikanth, V.; Kadry, S.; Taniar, D.; Damaševičius, R.; Rauf, H.T. Breast-cancer detection using thermal images with marine-predators-algorithm selected features. In Proceedings of the 2021 Seventh International Conference on Bio Signals, Images, and Instrumentation (ICBSII), Chennai, India, 25–27 March 2021; pp. 1–6.

[24] Hamed, G.; Marey, M.A.E.R.; Amin, S.E.S.; Tolba, M.F. Deep learning in breast cancer detection and classification. In The International Conference on Artificial Intelligence and Computer Vision; Springer: Cham, Switzerland, 2020; pp. 322–333.

[25] Mrs. Ritika Dhabliya. (2020). Obstacle Detection and Text Recognition for Visually Impaired Person Based on Raspberry Pi. International Journal of New Practices in Management and Engineering, 9(02), 01 - 07. https://doi.org/10.17762/ijnpme.v9i02.83

[26] Abdar, M.; Zomorodi-Moghadam, M.; Zhou, X.; Gururajan, R.; Tao, X.; Barua, P.D.; Gururajan, R. A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognit. Lett. 2020, 132, 123–131.

[27] Breast cancer dataset: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin +diagnostic

[28] Masciari, S.; Larsson, N.; Senz, J.; Boyd, N.; Kaurah, P.; Kandel, M.J.; Harris, L.N.; Pinheiro, H.C.; Troussard, A.; Miron, P.; et al. Germline E-cadherin mutations in familial lobular breast cancer. J. Med. Genet. 2007, 44, 726–731.

[29] Chaudhury, A.R.; Iyer, R.; Iychettira, K.K.; Sreedevi, A. Diagnosis of invasive ductal carcinoma using image processing techniques. In Proceedings of the 2011 International Conference on Image Information Processing, Shimla, India, 3–5 November 2011; pp. 1–6.

[30] Pervez, S.; Khan, H. Infiltrating ductal carcinoma breast with central necrosis closely mimicking ductal carcinoma in situ (comedo type): A case series. J. Med. Case Rep. 2007, 1, 83.

[31] Memis, A.; Ozdemir, N.; Parildar, M.; Ustun, E.E.; Erhan, Y. Mucinous (colloid) breast cancer: Mammographic and US features with histologic correlation. Eur. J. Radiol. 2000, 35, 39–43.