**International Journal of Innovative Technology and Exploring Engineering (IJITEE)**
**ISSN: 2278-3075 (Online), Volume-9 Issue-10, August 2020**

# Prediction and Classification into Benign and Malignant using the Clinical Testing Features

**Olawale Adepoju, Devaraj Verma C**

***Abstract:** Breast Cancer is the most often identified cancer among women and a major reason for the increased mortality rate among women. As the diagnosis of this disease manually takes long hours and the lesser availability of systems, there is a need to develop the automatic diagnosis system for early detection of cancer. The advanced engineering of natural image classification techniques and Artificial Intelligence methods has largely been used for the breast-image classification task. Data mining techniques contribute a lot to the development of such a system, Classification, and data mining methods are an effective way to classify data. For the classification of benign and malignant tumors, we have used classification techniques of machine learning in which the machine learns from the past data and can predict the category of new input. This study is a relative study on the implementation of models using Support Vector Machine (SVM), and Naïve Bayes on Breast cancer Wisconsin (Original) Data Set. With respect to the results of accuracy, precision, sensitivity, specificity, error rate, and f1 score, the efficiency of each algorithm is measured and compared. Our experiments have shown that SVM is the best for predictive analysis with an accuracy of 99.28% and naïve Bayes with an accuracy of 98.56%. It is inferred from this study that SVM is the well-suited algorithm for prediction.*

***Keywords:** Breast Cancer, Data Mining, Machine Learning, Naïve Bayes, Support Vector Machine.*

## I. INTRODUCTION

In India, the causes of cancer are virtually the same as elsewhere in the world. Unregulated, and unorganized proliferation of cells (carcinogens) is due to the chemicals, biological as well as other ecological attributes. In general, carcinogens interfere with normal cell DNA, leading to a variety of multi-step processes that are accountable for cancer metastasis or tumors under different circumstances. Cancer can be caused by internal causes such as genes, hormones, and immunes or by environmental factors such as cigarettes, nutrition, radiation, or other infectious agents. The initial or hypothetical symptom of breast cancer generally consists of a lump that is separate from the breast tissue surrounding it. Breast cancer can also be indicated by bumps found in the lymph nodes in the armpits. In order to determine the possibility of cancer by manual screening techniques, yet another test will be performed to establish whether a lump has been identified as a cancer screening.

In recent years, cancer of the breast is a significant reason for mortality in females. Whose lethal effect of the latter has contributed to prompt diagnosis of cancerous breasts by majority of countries worldwide, particularly speaking of developed countries, increasing the likelihood of succeeding in diagnosis [1]. It is therefore extremely important to concentrate efforts to develop an Integrative framework for breast cancer classification and detection [2], 23% of recurrent cancer in women has been confirmed to be breast cancer. In 2008, 40,000 people were put to death globally. Females between 40 and 69 years old are more at risk of infection [3]. The application of information extraction technology in the health sector can provide assistance to forecast decisions, reduce drugs prices, and help humanity to gain their health, increase the efficiency of health system, and give liberation to humanity. The categorization of cancer of the breast could bring about avoidance of unnecessary diagnosis of the clients. Therefore, studies would be initiated in other to detect categorize individuals across non-harmful and harmful tumor groups accordingly for accurate identification. With the development within the identification of essential or key features from diverse data sources, the intelligence of machine has been universally recognized. The implementation of machine classification technology can opt to be better used to classify cancerous and non-cancerous tumors. Within this scope, quite great deal of study is done utilizing different advance intelligence technologies on a large number of data of cancer of the breast. Many gives an outcome that classifying techniques provide good precision for forecasting the group of tumor.

Some many methods for detecting or diagnosing the disease have indeed been employed. Mammograms Sensitivity is estimated to be 70% to 90% with false negativity between 10 % and 30% of them. This occurs whenever the tests show that perhaps the region is clear when it is actually defective. In other words, the mammogram may neglect about a fourth of the tumors. This occurs often with dense breasts that make it hard to differentiate between masses [4]. In contrast, when the mammogram activates false positives, anomalous situations when the entire system is safe. Anomalous mammograms proceed to determine if the abnormal laity is cancerous, up with biopsy procedures.

---

Revised Manuscript Received on August 30, 2020.

* Correspondence Author

**Olawale Adepoju***, Department of Computer Science and Engineering, Jain (Deemed to be University), Bangalore, India. E-mail: olawale_adepoju@yahoo.com

**Dr. Devaraj Verma C**, Associate Professor, Department of Computer Science and Engineering, Jain (Deemed to be University), Bangalore, India. E-mail: c.devaraj@jainuniversity.ac.in

The consequence of false positive can be minimized by double testing and Mammography images. Two different physicians must check the same mammogram at different times. Although double tests demonstrated a 15% improvement in accuracy compared to single tests by the correct detection. It's a time consuming and costly process, though. The use of machine learning is a possible solution, algorithms for machine learning were proposed and implemented to help predict and classify breast cancer. The main purpose of this analysis is to develop the latest computerized interface Classification of the breast with cancer and non-cancer classification technique Correlating the performance of different classifiers utilized under methods regarding Accuracy rate, sensitivity, specificity, F1 score and improving the accuracy of prediction over existing solutions.

## II. RELATED WORKS

Categorization has been one of machine learning and data mining's most significant and basic practice. Much work has indeed been carried out to assess breast cancer by performing data analysis and machine learning on numerous sets of data, some of them are very precise in classification.

Comparing performance requirements for supervised learning performance, “Vikas Chaurasia and Saurabh Pal” [5], comparing between best classification for data on breast cancer, for classifiers “Naive Bayes, SVM-RBF Kernel, RBF Neural Networks, Decision trees (J48)”. Showing the results of the show that naive bayes is more efficient than any of the other classifiers executed on “Wisconsin breast cancer (original)” datasets which produces a predictive accuracy of 97.36 percent. The SVM classification technique was used by Alireza Osareh and Bita Shadgar [6] across two separate breast cancer test data-sets with 98.80% and 96.33% accuracy. The detection and forecasting of breast cancer recurrence are focused on by K-Nearest Neighbor, Support Vector Machine, Naïve Bayes, and Logistic Regression approaches, implemented in MATLAB, by “Mandeep Rana, Pooja Chandorkar and Alishiba Dsouza”. Two datasets from the UCI repository is considered for this classification. They are being used to classify the illness (WDBC), and the other is being used to forecast the recurrence (WPBC) [7]. The output of the Decision tree, Naïve Bayes, SVM, and KNN was compared by S. Aruna and L.V Nandakishore [8] to determine the best WBC classification. SVM indicates that it is the most reliable with an accuracy of 96.84%. Delen et al [9], used logit regression which is also an analytical method for developing models which are predictive using a large dataset and also used two major machine learning methods, “artificial neural networks, and decision trees with 10-fold cross-validation methods” have been utilized to calculate unbiased estimation for comparative analysis of the different forecasting. These outcomes showed decision tree to be the strongest indicator for the validation data set with 93.6% accuracy. WBCD data has been extracted by Naresh Khuriwal Nidhi Mishra and diagnosed with breast cancer. The outcomes of their studies showed the better functioning of ANN and Logistic Algorithms and gave a successful solution. It hit 98.50 percent precision [10].

### A. Data Mining

Mining of information is a process used in identifying useful information, trends, and patterns from raw data and it combines various areas of study, including machine intelligence, computer science, and analytics. More so analysis with manipulation of data are very necessary for the usage of the mined data to be interpreted. Classification is a data mining feature that attributes objects in a dataset for specific groups or categories. The aim of classifying deals with accurate prediction of the class labels respectively. A categorization task initiate with a data-set which identifies which group tasks are intended as targeted predictive variables. Binary classification is the least complex type of classification problem, the target variable only has two conceivable values in binary classification, for example, the presence of cancer or, on the other hand, the absence of breast cancer. The binary classifiers are therefore the most suitable algorithm for detecting and classifying breast cancer.

### B. Machine Learning

A collaborative field dealing with research which integrates insights into a spectrum of scientific fields, including machine intelligence, statistical analysis, and IT is referred to as machine learning. The benefit is that machine gains from previous or past experiences. The impetus behind machine learning algorithms is to learn "from the present knowledge," without explicitly been programmed. The machine learning approach allows users to uncover a hidden framework and build a separate predictive structure for large data sets. The target value is clearly described in the training process by the intelligent learning model, if trained, the prototype could be intended to foresee unknown target values in other data samples. Machine learning may be categorized as supervised or unsupervised subjected to if the data been trained are classed or labeled. Learning of machine focuses on discovering the ratio of inputted values to the outputted values relationship to forecast new outcomes if more inputs arise.

### C. Breast Anatomy

The basic understanding of breast and other breast dysfunction is important for understanding the significance of cancer of the breast diagnosis. Breast cancer can be said to be harmful tumor formed created by a cell division disorder. Mammography is a specific type of x-rays that identify any shape which suggests an underline disease, especially cancer, utilizing radiation levels between explicit intervals in an effort to obtain the breast images. Early diagnosis is highly critical on account of mammalian pathologies. The technical advancements in photography have helped to make cases of breast cancer more effective. For this domain, the identification of abnormalities in the early phases and the development of a good diagnosis are necessary for mammography. The epidermis causes tension in the area of the mammalian gland during the fetus phase to the upper ventral locale to the right and left sides.

Breast exists in both female and male, however, in females, the mammary glands generally evolved, with the exception of hormonal issues in some circumstances. The mamilla is also a small cone-shaped pinnacle enclosed with round zone consisting areola that consist enormous fatty organs which can be sometimes invisible, and skin with pigment.

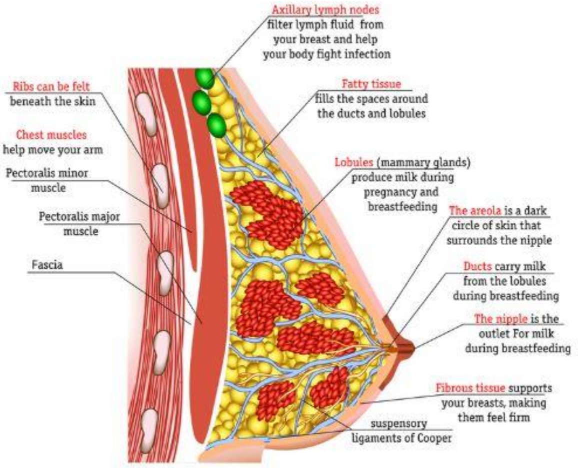


**Figure 1. Image of breast Anatomy**

At adolescence, the female breast typically develops as indicated by the growth of the gland and in turn of events increases the deposition of fat, likewise the nipple and areola develop. The size and state of the breast rely upon hereditary, ethnic, and nutritional elements. During the pregnancy, the color of the areola is dark, and then the pigmentation is preserved. The color reduces after breastfeeding has stopped, however, it is never completely lost during lifespan [11]. The breast comprises of the gland tissues, fibrous tissues, and gaps within lobes that attaches it to the fatty tissues. The breast comprises between 12 and 20 lobes of glandular tissue in each breast but varies in each woman that makes up the parenchyma of the mammary gland. Such lobes are very fatty in nature and consist of lobules bound to each other by areolar tissue, nerve cells, and ducts.

### D. Breast Cancer

The manner in which the cancer cells appear under the microscope will distinguish breast cancer into various types. Breast cancer is quite commonly carcinomas, which occurs in cells which fill tissues and organ like the breast. However, breast cancer is mostly a carcinoma category that begins with glandular tissues called adenocarcinoma.

There is no reliable method to avoid breast cancer, early detection is also the initial critical step in the treatment of breast cancer. It is critical in the diagnosis and treatment of breast cancer. Statistical data and facts on breast cancer educate us that evaluated new female cases and mortality is calculated by age.

#### Types of cancer tumor

Benign

This type of tumor is not cancerous, typically these can be removed and not necessarily develop back when they are out. When they are removed the tumor doesn't grow and it's rare for the benign tumor to be life-threatening.

**Figure 2. Image of benign tumor**

Malignant

Malignant tumors are carcinogenic. The cells are anomalous and segregate arbitrarily. The cells carry on forcefully and damage the tissue surrounding them. These can however pass off the malignant tumor to create new tumors in different areas of the body and into blood circulation. Numerous infections contaminate people yet just a couple of infections are known to advance cancer, these incorporate DNA infections, and retroviruses, a sort of RNA infection.

**Figure 3. Image of malignant tumor**

## III. METHODOLOGY

### A. Data Collection

Dataset used was retrieved from University of California, Irvine repository (UCI). It is WBCD (original) dataset, there are 699 cancer samples with 11 attributes.

The dataset consists of the following attribute information's: ID number, Clump thickness, Uniformity of cell size, Uniformity of cell shape, Marginal adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, Class (2 for Benign, 4 for Malignant).

### B. Data Preprocessing

In regards to data, it is of the belief that we are dealing with a dataset of different formats that can be structured, unstructured, log files. The machine cannot read images or categorical data, they can only read binaries, and so it did probably be good if our machine learning model gets trained by processed data. Our primary aim of this study is forecast and classify cancerous and non-cancerous breast from the medical test features given in the dataset. These features are used in the processing and for training the machine learning model. The techniques to identify the features and for each class of breast tumor can be seen in figure 4 below.

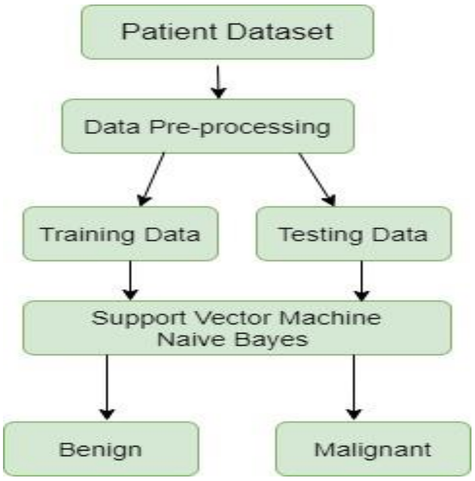


**Figure 4. A Systematic Diagram of Methodology**

Data processing is a necessary process in Machine Learning because the nature of information and the valuable data that can be gotten from it actually influences the model's capacity to learn, in this way, it is critical to prepare our information before it is incorporated into our model.

### i. Quality of Data Evaluation

After the data collection from the UCI repository, the first step of data preprocessing was to check the data quality assessment. Sometimes it is not applicable to all problems but dependent on the data for the study. In realistic data, there are indeed several null values to be expected due to human mistakes, measuring system errors or data gathering failures can be a concern. It doesn't generally make a difference if it's a classification or regression problem. No model can deal with or simulate such Null or NaN values by itself, hereby need human intervention.

Firstly, the total number of samples checked in the dataset which was 699 samples with 11 features or attributes and the first attribute which is the ID will be removed because it is not a feature we want to work within the classification model, and also the last attribute comprises a binary value where "2 represent benign tumor" and "4 represent malignant tumor" which is the class. The dataset contains 458 benign and 241 malignant as shown below.

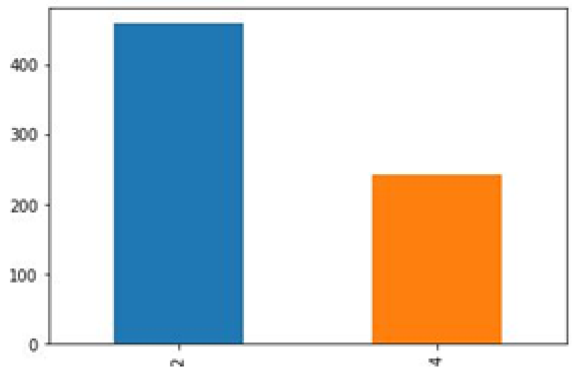


**Figure 5. Plot chart of "WBC (original)" dataset**

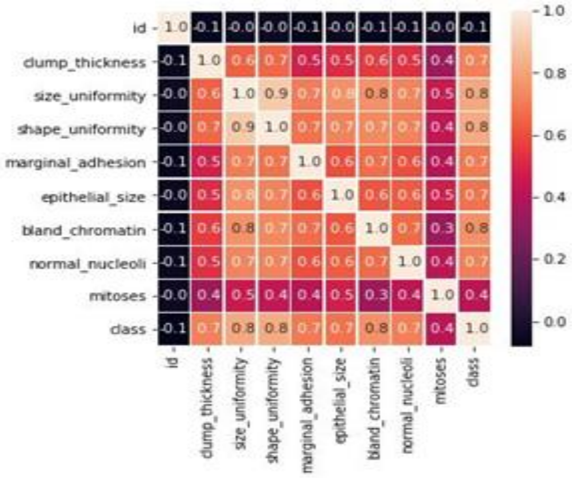
Also I checked if I had NaN values in the dataset, turned out that there were 16 missing value in the bare_ nucleoli column. Checking the null values in the dataset can be done using the isnull () method in python, it returns a Boolean matrix, if the value is NaN then True otherwise False. More so isnull () .sum () returns the column names alongside the quantity of NaN values in that specific segment.We can deal with this issue in different ways. The simplest approach to tackle this issue is by removing or dropping the null value columns or row using the dropna () function. Although it isn't the best choice to remove the null values from our dataset as it can prompt loss of significant data. In the event where you have substantial amount of null value, it can influence the accuracy of the dataset, which in reality large amount of null values happens to fill dataset.

In any case, the most basic strategy for managing missing values is by filling them in with the mean, median or mode estimation of individual features.

### ii. Principal component analysis using dimensional reduction

There are many features in almost every real-world datasets. For instance, we may need to manage a large number of features, likewise called dimensions. As the name implies, the reduction of dimension is aimed at reducing the number of features, however not only by choosing a sample of attributes from the list of features.

The reduction in dimension is employed to obtain 2-dimensional data with the goal to better represent models by plotting the area of the forecast and the boundaries of each model. Irrespective of the number of independent variables, we frequently wind up using an appropriate dimension reduction method with two independent variables. And some of the major advantages of dimensional reduction are based on that the data analysis algorithm function more best if the dimension of the dataset is lower, absolutely because the attributes of no value and noise are been eliminated. Also, models which are built on lower dimension data are more clear and comprehensible, thereby making data easier to visualize. Visualizing the data makes it easy to place the data in a visual context in order for the patterns, trends, and correlation which is not detected to be seen. Correlation often shows the relationship between two or more variables or features of a dataset.



| | id | dump_thickness | size_uniformity | shape_uniformity | marginal_adhesion | epithelial_size | bland_chromatin | normal_nucleoli | mitoses | class |
|---|---|---|---|---|---|---|---|---|---|---|
| id | 1.0 | -0.1 | -0.0 | -0.0 | -0.1 | -0.0 | -0.1 | -0.1 | -0.0 | -0.1 |
| dump_thickness | -0.1 | 1.0 | 0.6 | 0.7 | 0.5 | 0.5 | 0.6 | 0.5 | 0.4 | 0.7 |
| size_uniformity | -0.0 | 0.6 | 1.0 | 0.9 | 0.7 | 0.8 | 0.8 | 0.7 | 0.5 | 0.8 |
| shape_uniformity | -0.0 | 0.7 | 0.9 | 1.0 | 0.7 | 0.7 | 0.7 | 0.7 | 0.4 | 0.8 |
| marginal_adhesion | -0.1 | 0.5 | 0.7 | 0.7 | 1.0 | 0.6 | 0.7 | 0.6 | 0.4 | 0.7 |
| epithelial_size | -0.0 | 0.5 | 0.8 | 0.7 | 0.6 | 1.0 | 0.6 | 0.6 | 0.5 | 0.7 |
| bland_chromatin | -0.1 | 0.6 | 0.8 | 0.7 | 0.7 | 0.6 | 1.0 | 0.7 | 0.3 | 0.8 |
| normal_nucleoli | -0.1 | 0.5 | 0.7 | 0.7 | 0.6 | 0.6 | 0.7 | 1.0 | 0.4 | 0.7 |
| mitoses | -0.0 | 0.4 | 0.5 | 0.4 | 0.4 | 0.5 | 0.3 | 0.4 | 1.0 | 0.4 |
| class | -0.1 | 0.7 | 0.8 | 0.8 | 0.7 | 0.7 | 0.8 | 0.7 | 0.4 | 1.0 |

**Figure 6.Visual correlation between features using heat map**

### A. Model Selection

The choice of algorithms is perhaps the most interesting stage in building a machine learning model. In large sets of data, one can implement more than one form of the machine learning algorithm. However, all the different algorithms can indeed be categorized into two-part as a substantial class.

The supervised learning approach is used to train the model on data with clearly labeled input and output. This model will learn from the input data which are also training data and analyze the prospective information in order to forecast outcomes.

These techniques are divided into classification and regression methods. The classification task is whether the outcome of prediction falls into a discontinuous form such as “benign” or “malignant”, “1” or “0”, “yes” or “no”. While regression task outcome is a continuous or real value such as “height”, “stock values”.

The unsupervised learning provides non-grouped and unlabeled data to the model, and it allows the algorithm to evaluate the information even without any instruction. The machine has giving train from data that are not labeled nor categorized to make the algorithm function without basic instructions. Having the dependent variable in our dataset, which is “2 for benign” and “4 for malignant”. Therefore supervised learning classification method is adopted, the selected algorithms are support vector machine (SVM), and Naïve Bayes.

#### i. Introduction to support vector machine

SVM is a grouping of supervised learning techniques, which are used mainly for highlighting machine learning groups and recurrent, SVM classifiers have the reduced features obtained during the PCA stage. At the output, the classifier forecast that the attributes compared to the labels are either cancer or non-cancerous. SVM is used as a non-probabilistic sampling classifier and generates a hyper-plane separating the two groups. The essential idea of driving Support Vector Machine (SVM) is a hyper-plane used in the decision. By plotting any acquired data value in an n-dimensional space or graph, a support vector algorithm is performed. The total number of data features present here is "n." The value of each data is displayed as a different graph coordinate. SVM is an example of supervised learning techniques that can be implemented in problems of classification and regression. Practically, the grouping procedure isn't straightforward, and, when it depends on the partition utilizing diverse lines it is known as hyper-plane classifiers. The optimum hyper-plane is chosen according to the line-distance. SVM Vector help divides the class from one side to the other. The difference is called the margin and the margin is referred to as vectors of support as illustrated in figure 7.

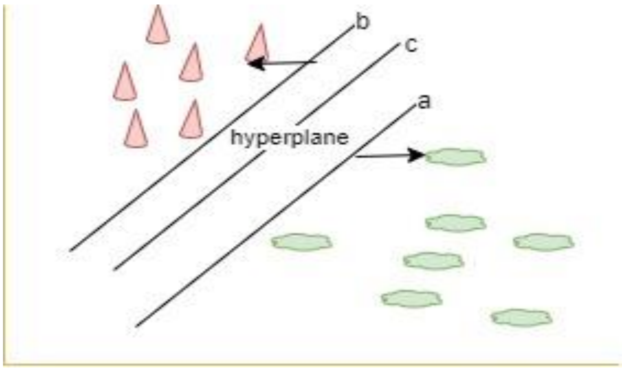


**Figure 7. A support vector machine hyper plane**

#### ii. Introduction to naïve bayes

The Bayesian classification model is a probabilistic classifier machine learning technique. The Bayes classification is based on theorem, the maximum option depending on chance, values and known likelihoods is predicted.

The best hypotheses (j) and data given (k) in machine learning is often chosen, the hypotheses (j) can represent class given to a present information (k) in a classification task.

Naïve bayes is represented by:

$$P(j|k) = \frac{\left(P(k|j)*P(j)\right)}{P(k)}$$

$P(j|k)$= likelihood of hypothesis (j) specified the data (k). It can be referred to as posterior probability.

$P(k|j)$= likelihood of data (k) stated hypothesis (j) is true.

$P(j)$= likelihood of hypothesis (j) is true irrespective of the data. It can be referred to as the prior likelihood of j.

$P(k)$= likelihood of data irrespective of the hypothesis.

The Bayesian theory gives a formula for the posterior likelihood calculation $P(j|k)$, the likelihood of results (j) given for certain conditions (k).

In order to link it to the previous likelihood of the outcome lacking awareness of the influence factors, the equation estimates the future likelihood by using a probability ratio $P(k|j) = P(k)$.

The proposition with maximum likelihood after determining the posterior probability for a variety of distinct assumptions will be chosen. This theory of naive bayes is built on the presumption that every factor independently influences the result and is therefore naive.

## IV. RESULT AND DISCUSSION

### A. Performance evaluation criteria for classification

The evaluation of performance for the machine learning model depends on a variety of variables like the classification accuracy, precision score, recall score, error rate, and f1 score. Testing and evaluating this machine learning model, the widely known procedure used in statistical classification was used called the confusion matrix.

The confusion matrix table gives an output that describes the overall performance of the model, it gives the visual presentation of the algorithm's performance. The number of correct and incorrect predictions are known.

**Table I: Confusion Matrix Table**

| | ***"Predicted No"*** | ***"Predicted Yes"*** |
|---|---|---|
| ***"Actual No"*** | "TN" | "FP" |
| ***"Actual Yes"*** | "FN" | "TP" |

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

Sensitivity (recall) = $\frac{TP}{TP+FN}$

Specificity = $\frac{TN}{TN+FP}$

Precision= $\frac{TP}{TP+FP}$

Error rate = $\frac{FP+FN}{TP+TN+FP+FN}$

F-measure (F1 score) = $\frac{(2*Recall*Precision)}{Recall+Precision}$

## B. SVM classifier performance

The parameters were analyzed by using SVM with linear kernel, confusion matrix provides the performance evaluation as shown in table 2.

**Table II: SVM Confusion Matrix Table**

| | ***"Predicted No"*** | ***"Predicted Yes"*** |
|---|---|---|
| ***"Actual No"*** | 103 | 0 |
| ***"Actual Yes"*** | 1 | 35 |

Accuracy = $\frac{103+35}{103+35+0+1} * 100 = 99.28\%$

Recall = $\frac{35}{35+1} * 100 = 97.22\%$

Specificity = $\frac{103}{103+0} * 100 = 100\%$

Precision = $\frac{35}{35+0} * 100 = 100\%$

Error Rate = $\frac{0+1}{103+35+0+1} * 100 = 0.72\%$

F1 Score = $\frac{2*97.22*100}{97.22+100} * 100 = 98.59\%$

## C. Naïve Bayes classifier

The parameters were analyzed by using naïve bayes, confusion matrix provides the performance evaluation as shown in table 3.

**Table III: Naïve Bayes Confusion Matrix Table**

| | ***"Pred No"*** | ***"Pred Yes"*** |
|---|---|---|
| ***"Actual No"*** | 102 | 0 |
| ***"Actual Yes"*** | 2 | 35 |

Accuracy (ACC) = $\frac{102+35}{102+35+0+2} * 100 = 98.56\%$

Recall (RC) = $\frac{35}{35+2} * 100 = 94.59\%$

Specificity (SP) = $\frac{102}{102+0} * 100 = 100\%$

Precision (PR) = $\frac{35}{35+0} * 100 = 100\%$

Error Rate (ER) = $\frac{0+2}{102+35+0+2} * 100 = 1.44\%$

F1 Score (F1) = $\frac{2*94.59*100}{94.59+100} * 100 = 97.22\%$

## D. Comparison of classifier performance

Considering the result of Accuracy score, sensitivity, specificity, precision, error rate, f-measure for various machine learning techniques is illustrated in table 4 below.

**Table IV: Comparison table of classifier performance**

| | ***Accuracy*** | ***Recall*** | ***Specificity*** | ***Precision*** | ***Error Rate*** | ***F1 Score*** |
|---|---|---|---|---|---|---|
| ***SVM*** | 99.28 | 97.22 | 100 | 100 | 0.72 | 98.59 |
| ***NB*** | 98.56 | 94.59 | 100 | 100 | 1.44 | 97.22 |

The table above shows the efficiency of the two models. The support vector machine technique has shown the greater and highest accuracy outcome in the test measures used. In contrast with the naïve bayes algorithm, it is seen that SVM is the most appropriate algorithm to predict the breast cancer data.

## V. CONCLUSION

The study made use of 699 data which present benign and malignant classes with 9 features, also with some missing data which has been replaced using backfill and accounted for the improvement in accuracy of the model. The SVM classifier modeled at an accuracy of 99.28% with an error rate of 0.72% and the naïve Bayes modeled at an accuracy of 98.56% with an error rate of 1.44%. In regards to the accuracy, recall, specificity, error rate, precision, and f1 score, the efficiency, and viability of the algorithms were compared. The SVM model performed better than the naïve Bayes in this study.

## ACKNOWLEDGEMENTS

The authors are grateful to Jain University, Department of Computer Science and Engineering, for their in-valuable support and feedback on this work.

## REFERENCES

1. Acha, B., Rangayann R.M, Desautels J.E. L. (2006). "Detection of micro calcifications in mammograms" .SPIE, Bellingham, Recent Advances in Breast Imaging, Mammography, and Computer Aided Diagnosis of Breast Cancer.
2. Choudhari G, Swain D, Thakur D, Somase, K. (2012).”An adaptive approach to classify and detect the breast cancer using image processing", International Journal of Computer Applications (0975-8887) 45(17).
3. Dheeba J. WiselinJiji G. (2010) “Detection of micro calcification cluster in mammograms using neural network “, International Journal of Advanced Science and Technology.
4. Erickson, Carissa (2005). “Automated detection of breast cancer using saxs data and wavelet features", (Unpublished doctoral dissertation) university of Saskatchewan, Saskatoon.
5. V. Chaurasia and S. Pal, “Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability,” vol. 3, no. 1, pp. 10–22, 2014.
6. AlirezaOsarech, BitaShadgar,”A Computer Aided Diagnosis System for Breast Cancer”, International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011
7. Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza, “Breast cancer diagnosis and recurrence prediction using machine learning techniques”, International Journal of Research in Engineering and Technology Volume 04, Issue 04, April 2015.
8. S. Aruna and L. V Nandakishore, “KNOWLEDGE B ASED ANALYSIS OF VARIOUS STATISTICAL TOOLS IN DETECTING BREAST” pp. 37–45, 2011.
9. D. Delen, G. Walker, and A. Kadam, “Predicting breast cancer survivability: A comparison of three data mining methods” Artif.Intell. Med., vol. 34, pp. 113–127, 2005. doi:10.1016/j.artmed.2004.07.002
10. N. Khuriwal and N. Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," 2018 IEEMA Engineer Infinite Conference (eTechNxT), New Delhi, 2018, pp. 1-5, doi: 10.1109/ETECHNXT.2018.8385355.
11. Moore K. L, Agur A.M and Dalley A.F (2004). Essential Clinical Anatomy. (4nd Ed.). (W. Kluwer, Ed.)

## AUTHORS PROFILE

**Olawale Adepoju,** Dept. of Computer Science and Engineering, Jain deemed to be University, Bangalore, India. He received Bachelor of Engineering from Ekiti State University (B.Eng.), M.Tech from Jain deemed to be University Bangalore, India. His research areas are Data Mining, Cloud Computing and Machine Learning.

**Devaraj Verma C**, Associate Professor, Department of Computer Science and Engineering, Jain (Deemed to be University), He is NET and KSET qualified professor. He is associated as a life time member of professional bodies like ISTE, IACSIT and IAENG. He had obtained his Ph.D. degree in the area of artificial intelligence. His PhD work is towards building the Computational model for diversity of Mind. He applied the theory of mind to detect the malicious node using the cognitive model. His area of interest includes cognitive science, image processing, IoT and machine learning in the area of Artificial intelligence. He is the principal investigator of the in-house project Automation of agriculture using IoT. He has published more than 20 research papers in various national/international conferences and journals. He is the reviewer of the book titled “*Artificial Intelligence: Theories and Applications*”. It is a 600-odd page book written by Dr. Lavika Goel, 2nd edition, published by Wiley India Pvt. Ltd, July 2019.