Machine Learning-Based Breast Cancer Prediction: Advancements and Insights for Early Intervention

Sameen Sadman

Thesis report submitted in partial fulfillment of the

requirements for the award of the degree of

MASTER OF DATA SCIENCE

CHARLES DARWIN UNIVERSITY

COLLEGE OF ENGINEERING, INFORMATION TECHNOLOGY AND ENVIRONMENT

June 2024

DECLARATION

I hereby declare that the work herein, now submitted as [an interim report/a report/a thesis report] for the degree of [Name of Degree] ([Specialisation, if applicable]) at Charles Darwin University, is the result of my own investigations, and all references to ideas and work of other researchers have been specifically acknowledged. I hereby certify that the work embodied in this [interim report/report/thesis report] has not already been accepted in substance for any degree, and is not being currently submitted in candidature for any other degree.

Signature: Sameen

Date: 2nd July 2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

RF: Random Forest

GBM: Gradient Boosting Machine

ROC: Receiver Operating Characteristics

AUC: Area Under Curve

SVM: Support Vector Machine

ML: Machine Learning.

AI: Artificial Intelligence.

LIST OF SYMBOLS

Machine Learning-Based Breast Cancer Prediction: Advancements and Insights for Early Intervention

Abstract— **Timely detection of breast cancer is essential, for improving outcomes by facilitating intervention. This research delves into the application of Machine Learning (ML) techniques like Support Vector Machines (SVM) k Nearest Neighbors (KNN) and logistic regression in the realm of breast cancer identification.**

By combining machine learning algorithms with advanced image analysis methods the automation of interpreting mammography images becomes possible. This approach has the potential to enhance both the precision and efficiency of image interpretation processes. The primary aim of this study is to develop a machine-learning system for accurately categorizing breast cancer in medical images. The key purpose of this system is to aid radiologists in identifying breast cancer cases thereby reducing errors.

The research encompasses gathering. Preprocessing a dataset of mammograms at an early stage followed by deploying and evaluating various machine learning models. We have used Random forest, KNeighbours Classifier and Naïve Bayes Classifier for our classification task. Accuracy we achieved are 96.49%, 95.61% and 97.36% respectively for Random forest, KNeighbours Classifier and Naïve Bayes Classifier.

Keywords— breast cancer, cancer cell, detection, Malignant and Benign, ML algorithm

I. INTRODUCTION

Breast cancer, a disease that predominantly impacts women plays a role, in the mortality rate of women worldwide. In 2020, 685,000 deaths across the globe were linked to this illness[2].

Early detection of breast cancer is vital in improving treatment outcomes and boosting survival rates.
Different diagnostic procedures such as MRI scans, mammograms, ultrasounds and tissue biopsies are crucial for identifying cells.
Our study utilizes a dataset comprising digitized images from FNA biopsies of breast masses to examine the characteristics of cell nuclei.

A key aspect in diagnosing breast cancer involves distinguishing between tumors that pose a risk.

Regrettably, not all healthcare professionals excel at differentiating between these types of tumors leading to a time consuming process that can last up to two days.
Machine learning algorithms hold promise in predicting the nature of cells.

Within the realm of intelligence (AI) machine learning empowers systems to learn and improve themselves based on data without programming. This paradigm shift enables computer programs to autonomously glean insights from data and expand their knowledge base.

Our study delves into machine learning algorithms such, as Support Vector Machine (SVM) Decision Tree (CART) Naive Bayes (NB) and k Nearest Neighbors (k NN).
We examined how breast cancer cells are categorized and evaluated the efficiency of these techniques by considering measures such, as correctness, exactness, retrieval and overall performance score.

One notable algorithm we analyzed is the Support Vector Machine (SVM). SVM is a machine learning tool that excels in handling classification and regression tasks with a focus, on classification. In SVM each data point is depicted as a coordinate in a space where 'n' represents the number of features. The goal is to discover a hyperplane that effectively divides the two classes. Support vectors, which represent coordinates of observations assist SVM in establishing this boundary by optimizing the margin between the decision hyperplane and neighboring instances. This leads to the creation of a classifier with generalizability that allows for classifying samples.

Moreover SVMs can also provide results that enhance the process. In the following sections we will delve into how SVM and other machine learning techniquesre used in categorizing breast tumors. Our research focuses on categorizing breast cancer cells to predict their occurrence than treating the disease itself. The identified hyperplane plays a role, as a decision boundary that helps in identifying and rectifying misclassifications thereby improving breast cancer diagnosis.

Background

Breast cancer is the most common cancer in women and the second leading cause of cancer death in the U.S. Originating from breast tissue, it accounted for 519,000 global deaths in 2004[1]. Cancer cells arise from mutations in DNA or RNA, often undetected by the immune system due to their similarity to normal cells. Mutations can occur spontaneously or be triggered by factors like radiation, viruses, chemicals, and aging. Cancer is termed an "Entropic Disease" due to increased entropy. It develops when the immune system fails or is overwhelmed, often influenced by unhealthy environments, poor diets, genetic factors, and advanced age.

Types of Breast Cancer:

- Infiltrating Lobular Carcinoma (ILC): Starts in milk glands, spreading to other body regions, comprising 10-15% of cases.
- Infiltrating Ductal Carcinoma (IDC): Begins in milk ducts, invades fatty tissue, and is the most common, at 80% of cases.
- Medullary Carcinoma: An invasive cancer with distinct tumor boundaries, accounting for 5% of cases.
- Mucinous Carcinoma: Rare, mucus-producing cancer with a better prognosis, forming a small percentage of cases.
- Tubular Carcinoma: Invasive with a better prognosis, making up about 2% of diagnoses.
- Inflammatory Breast Cancer: Fast-growing, rare (1%), causing inflamed, dimpled skin.
- Paget's Disease of the Nipple: Rare (1%), spreading from milk ducts to the nipple and areola skin.

Aim of Research

The aim of this research is to detect breast cancer as early as possible with highest precison. The main purpose of this research paper can be broken down to:

- Apply data preprocessing techniques before implementation of the proposed algorithms.
- Implementation of four different Machine Learning models to classify breast cancer.
- Compare and evaluate the results of different models based on Accuracy, Precision, Recall and f-1 score.
- Reduce the reliance on manual detection of breast cancer which is time consuming and also expensive.
- Find out the reasons behind breast cancer and predict them as accurately as possible.

Structure of Paper

In this research we have conducted classification of breast cancer upon two classes: benign and malignant. This paper includes:

- Dataset
- Preprocessing
- Approach
- Experimental Results
- Conclusion

II. APPROACH

The methodology employed in this study is a straightforward one that includes gathering the data, cleaning it, pre-processing it, visualizing it, and analysis, followed by the application of algorithms to create a model that categorizes fetal health. The following steps outline the approach:

Dataset and Attributes:
The dataset contains two classes- 357 benign and 212 malignant. The attributes include real valued features for each cell nucleus such as- radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The dataset does not contain any missing attributes and all the feature values have been recoded with four significant digits. All the features form a digitized image of a fine needle aspirate (FNA) of a breast mass.

Data Analysis:

The researchers carefully analyzed the data collected from the reviewed papers to identify themes, emerging trends and gaps, in the research. They then conducted a analysis to summarize key statistics such, as the types of machine learning techniques used characteristics of datasets and accuracy of predictive models. This numerical perspective helped to gain an understanding of how different machine learning approachesre utilized in predicting breast cancer.

Preprocessing: To make sure the dataset is ready for analysis, a number of data preparation techniques are used before the model is implemented. This covers encoding categorical variables, managing missing values, and normalizing features. To maximize the prediction performance of the models, feature selection techniques are also used to find the most pertinent predictors.

Algorithm Selection:

✓ Random Forest(RF):
Builds a reliable and accurate classifier that can handle big datasets with high dimensionality by using an ensemble of decision trees. Overfitting is lessened by Random Forest, which aggregates forecasts from many decision trees.

✓ Support Vector Machine(SVM):
SVM is a potent technique for classification tasks that seeks to locate the ideal hyperplane in the feature space that maximum divides classes in order to achieve high classification precision. When data cannot be separated linearly, SVM is specially useful.

✓ Gradient Boosting Machine(GBM):
By reducing loss functions, GBM iteratively

constructs a sequence of decision trees, each of which corrects the mistakes of the one before it, resulting in improved predicting performance. Complex interactions and nonlinear correlations within the data are easily captured by GBM.

✓ Naïve Bayes:
Naive Bayes is a computationally efficient and effective method for classification jobs since it relies on the Bayes theorem and computes the probability of class membership for a given occurrence assuming predictor independence. Large feature datasets benefit greatly from the application of Naive Bayes.

✓ K Nearest Neighbour(KNN):
KNN is an instance-based, non-parametric method that groups instances in the feature space according to how similar they are to each other. By designating the majority class label among its k closest neighbors, KNN generates predictions. It works especially effectively with datasets that include distinct clusters and local patterns.

These algorithms were selected because they can work with different kinds of data and have a good chance of correctly identifying cases of breast cancer.

III. Experiments / Testing / Measurements / Data

Naïve Bayes:

The statistical Bayes theorem forms the basis of Naive Bayes. It predicts by calculating the probability, for each class under both unconditional assumptions. In classification we aim to determine the probability of a class label (C) being assigned based on the features (X). The naive Bayes classifier assumes that a features presence, in a class is independent of any features existence the term "naive" assumption. The likelihood of a class label given the features can be calculated using the Bayes theorem in this manner:

P(C|X)=$\frac{P(C)*PX|C)}{P(X)}$

Where:

- $P(C|X)$ is the posterior probability of class CC given the feature vector XX.
- $P(C)P(C)$ is the prior probability of class CC, which is the proportion of instances of class CC in the training data.
- $P(X|C)P(X|C)$ is the likelihood of the feature vector XX given class CC.
- $P(X)P(X)$ is the marginal likelihood or the evidence, which is the total probability of the feature vector XX occurring under all possible classes.

- P(C|X) is the posterior probability of class C given the features X.
- P(X|C) is the probability of X given class C.
- Class C's prior probability is denoted by P(C).
  P(X) represents the prior probability of feature X.

Gradient Boosting Machine:

For regression and classification applications, the potent ensemble learning method known as Gradient Boosting Machine (GBM) is employed. It creates models one after the other, trying to fix the flaws of the models that came before it. Through the process of integrating the strengths of numerous weak learners, usually decision trees, a highly accurate prediction model is produced.

How Gradient Boosting Works:
Initialization:

The process starts with an initial model, often a simple one like the mean of the target values (for regression) or a model that predicts a constant value.
Calculate Residuals:

The residuals are the differences between the actual values and the predictions of the current model.
These residuals represent the errors that the next model needs to address.
Fit a New Model:

A new model is trained on the residuals of the previous model.
This new model tries to predict the errors of the current ensemble model.
Update the Ensemble:

The predictions of the new model are added to the ensemble of previous models.
The update rule usually involves a learning rate that controls the contribution of each new model.
Repeat:

Steps 2-4 are repeated for a predefined number of iterations or until the model performance stops improving on a validation set.

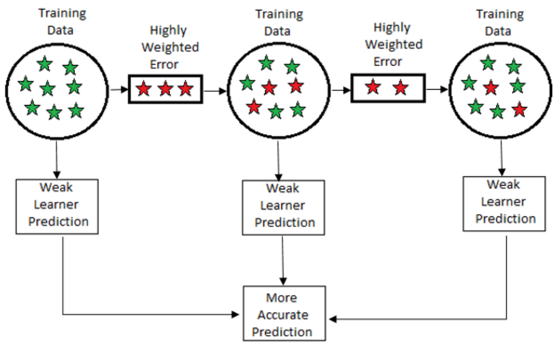


Fig1 : GBM Architecture[3].

KNN:

A straightforward instance-based learning approach called k-Nearest Neighbors (k-NN) is used to both regression and classification problems. It functions according to the idea that comparable instances are located close to one another in feature space. This is a thorough explanation of the functions, benefits, drawbacks, and uses of k-NN.

Training Phase:

k-NN is a type of lazy learning, meaning it does not involve an explicit training phase. The algorithm simply stores the entire training dataset.

Prediction Phase:

Classification: For a new instance, k-NN finds the k training instances closest to the new instance based on a chosen distance metric. The class label is then assigned based on the majority class among these k neighbors.

Regression: For regression, the algorithm predicts the target value as the average (or weighted average) of the target values of the k nearest neighbors.

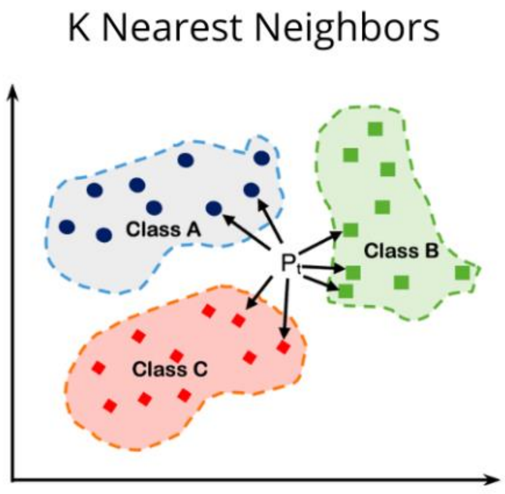


Fig2 : KNN Architecture[4].

Random Forest:

Leo Breiman and Adele Cutler are the trademark holders of the popular machine learning technique known as "random forest," which aggregates the output of several decision trees to produce a single outcome. Its versatility and ease of use, combined with its ability to handle both regression and classification issues, have driven its popularity.

The three primary hyperparameters of random forest algorithms must be specified before to training. These consist of the size of the nodes, the count of trees, and the quantity of characteristics sampled. Regression and classification issues can then be resolved using the random forest classifier.

Each decision tree in the ensemble of decision trees used in the random forest technique is made up of a bootstrap sample, which is a sample of data taken from a training set with replacement. One-third of the training sample is designated as test data; this is referred to as the out-of-bag (oob) sample, and it is something we will discuss more.

Feature bagging is then used to provide even further randomization, increasing dataset variety and decreasing decision tree correlation. The prediction's decision will change depending on the kind of difficulty. The individual decision trees in a regression job will be averaged, and in a classification work, the predicted class will be determined by a majority vote, or the most common categorical variable. Lastly, cross-validation is performed using the sample to complete that prediction.

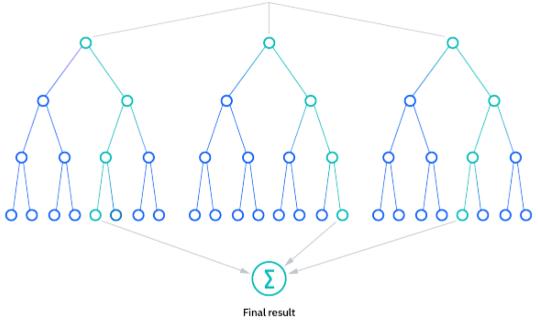


Fig3 : Random Forest Calssifier[5].

Support Vector Machine:

Support Vector Machine (SVM) is a powerful and versatile supervised machine learning algorithm used for both classification and regression tasks, though it is more commonly used for classification.

The advantages of support vector machines are:

Effective in high dimensional spaces.

Still effective in cases where number of dimensions is greater than the number of samples.

Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.

SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation

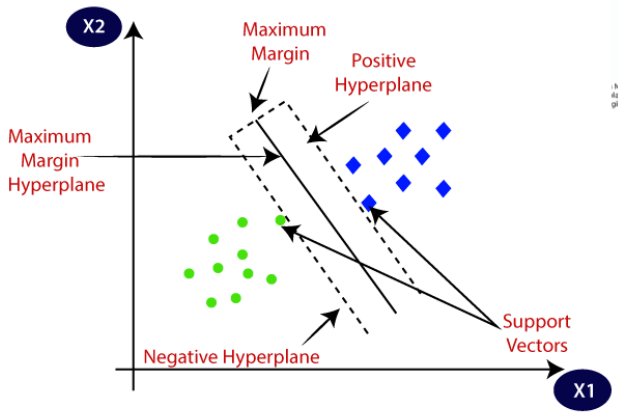


Fig4: How SVM Works[6].

IV. Analysis and Discussion of Results

This section describes the output and results of different algorithms on our dataset. We have included Classification Report, Confusion Matrix and ROC_AUC curve for each model on our dataset.

1.Random Forest:

Random Forest Classifier worked very well on our dataset. We achieved overall 96.49% accuracy with this model with precision of 97% , recall of 96% and f-1 score of 96%.

With the following figures we can evaluate our model for the given breast cancer dataset which included two classes of benign and malignant(class 0 and 1 on figures).

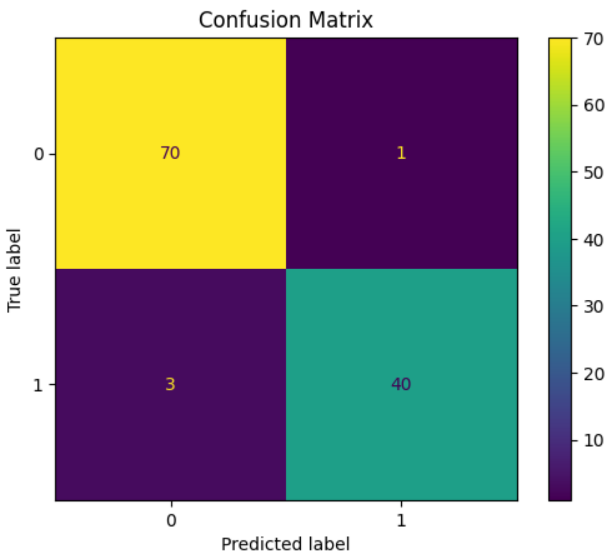


Fig5 : Confusion Matrix for RF.

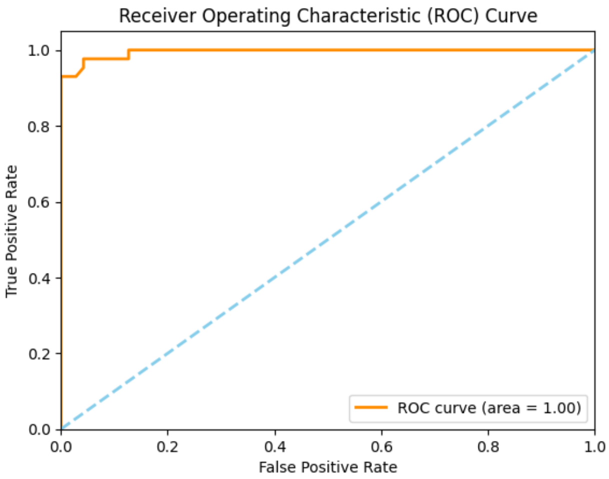
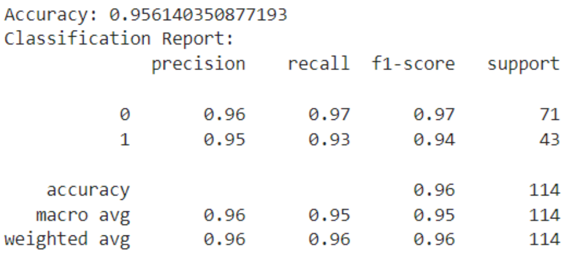


Fig6 : ROC_AUC Curve for RF.



```
Accuracy: 0.956140350877193
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.97      0.97        71
           1       0.95      0.93      0.94        43

    accuracy                           0.96       114
   macro avg       0.96      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114
```

Fig7: Classification Report for RF.

2.Support Vector Machine:

While using svm we achieved an overall accuracy of 96%. The following figures describe how well the algorithm worked for the breast cancer dataset.

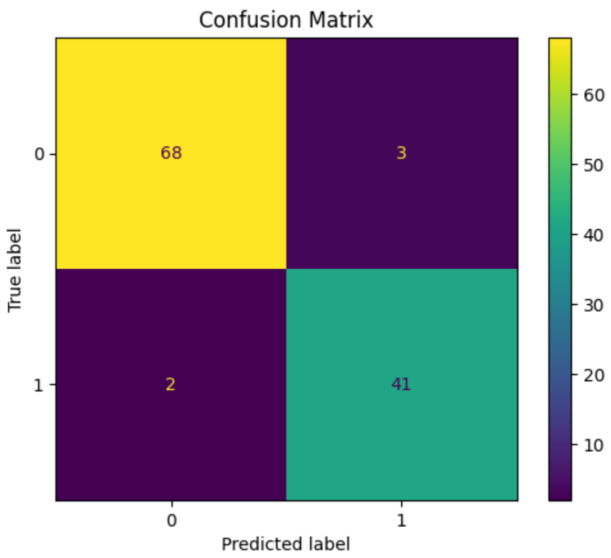


Fig8: Confusion Matrix for SVM.

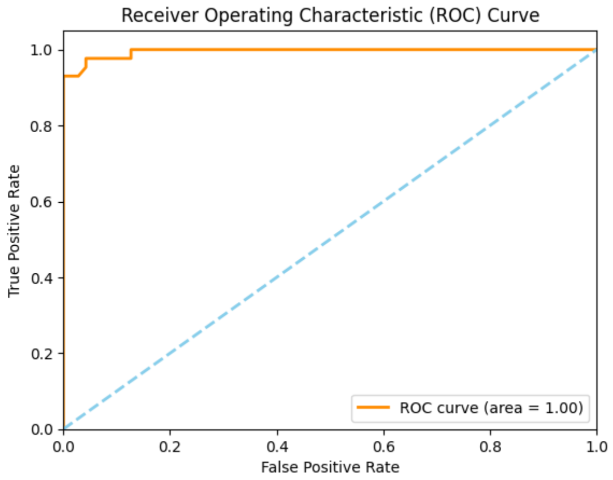
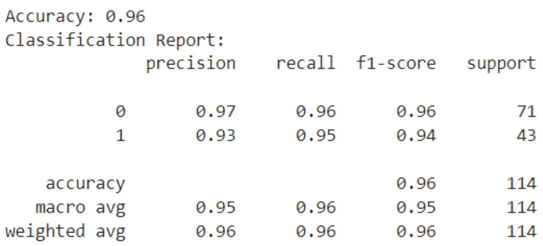


Fig9: ROC_AUC curve for SVM.



```
Accuracy: 0.96
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.96      0.96        71
           1       0.93      0.95      0.94        43

    accuracy                           0.96       114
   macro avg       0.95      0.96      0.95       114
weighted avg       0.96      0.96      0.96       114
```

Fig10: Classification Report for SVM.

3. Naïve Bayes:

Using Naïve bayes we achieved an overall accuracy of 97.36% which was the best among all the algorithms that we have used.

The following figures with Confusion Matrix ,ROC_AUC curve and classification report shows how good it was for the classification task.

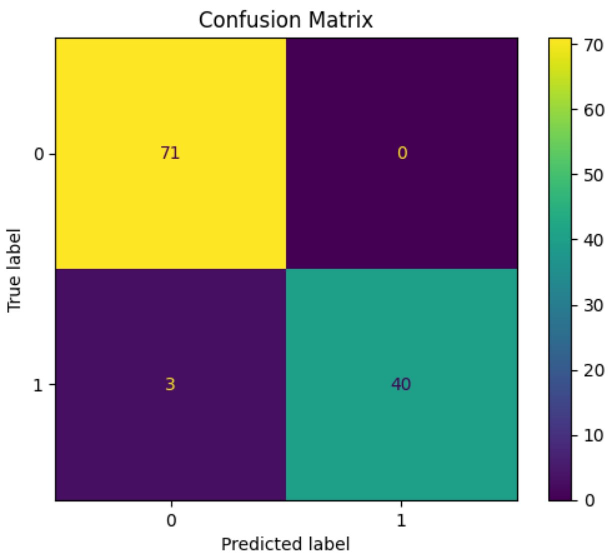


Fig11: Confusion Matrix for Naïve Bayes.

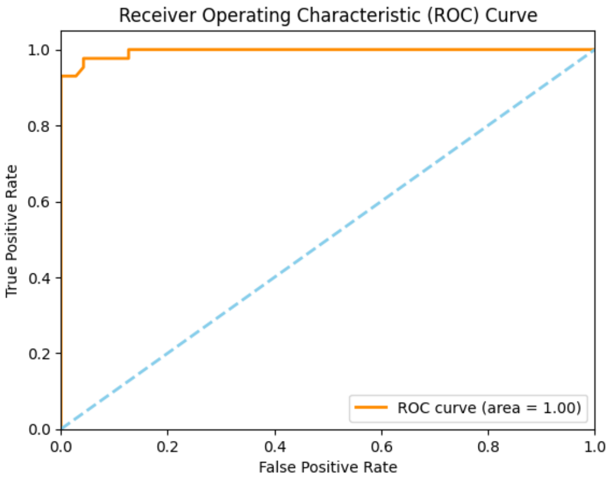
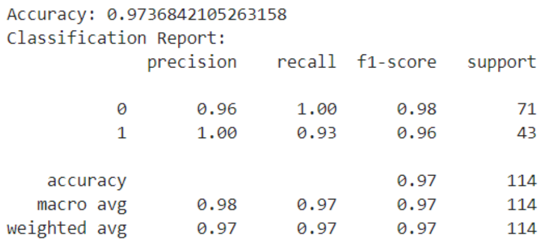


Fig12: ROC_AUC Curve for Naïve Bayes.



```
Accuracy: 0.9736842105263158
Classification Report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        71
           1       1.00      0.93      0.96        43

    accuracy                           0.97       114
   macro avg       0.98      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114
```

Fig13: Classification Report of Naïve Bayes.

4.Gradient Boosting Machine(GBM):

We achieved an overall accuracy of 95.61% with GBM on our dataset.

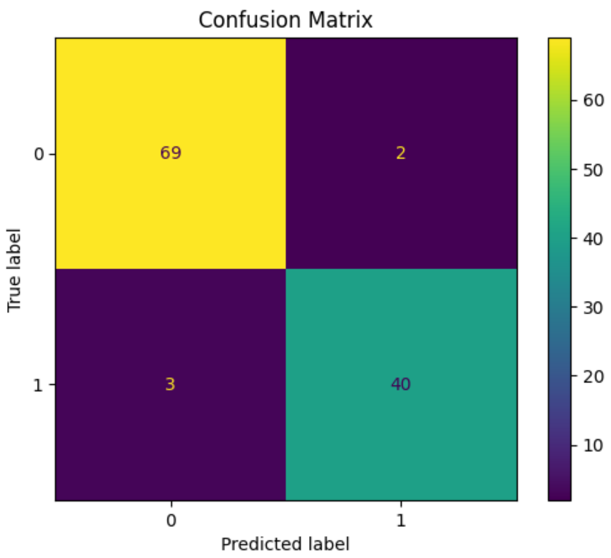


Fig14: Confusion Matrix for GBM.

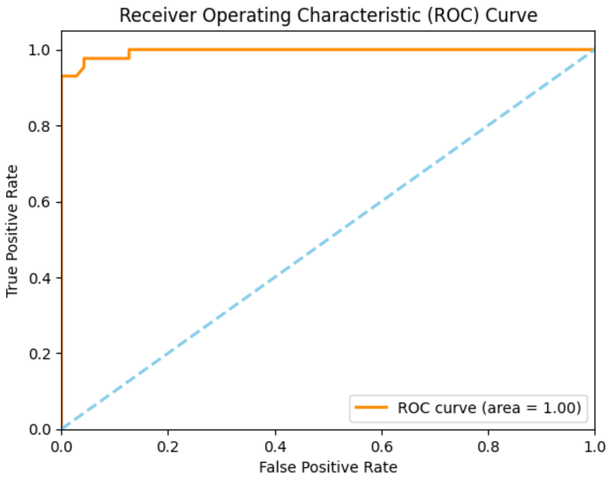
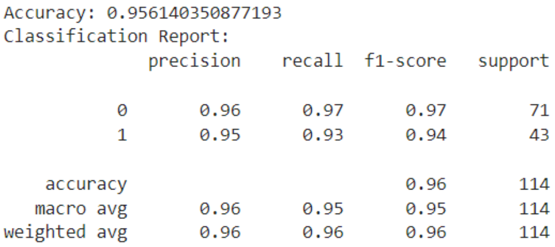


Fig15: ROC_AUC Curve for GBM.



```
Accuracy: 0.956140350877193
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.97      0.97        71
           1       0.95      0.93      0.94        43

    accuracy                           0.96       114
   macro avg       0.96      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114
```

Fig16: Classification Report for GBM.

5.KNN:

For KNN we got identical accuracy and results with GBM. The results were the same with these two algorithms. The figures that describe the match are given below.

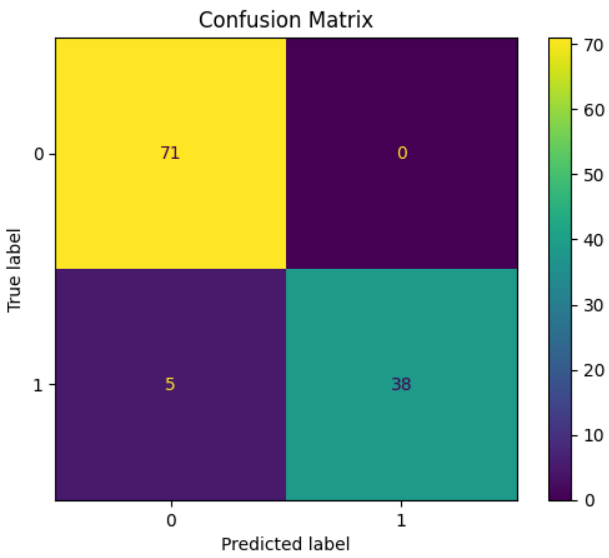


FIG 17: Confusion Matrix for KNN.

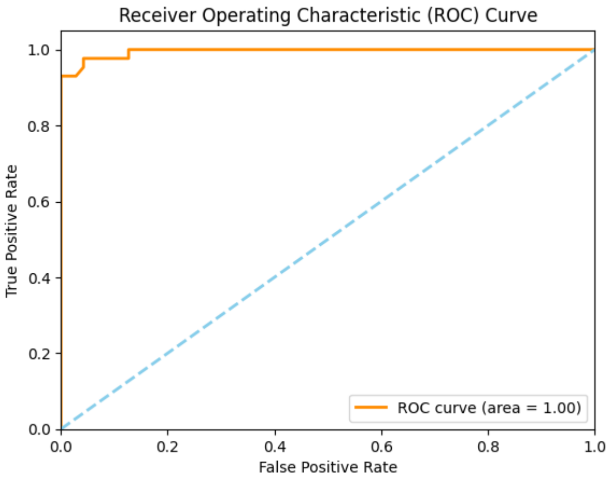
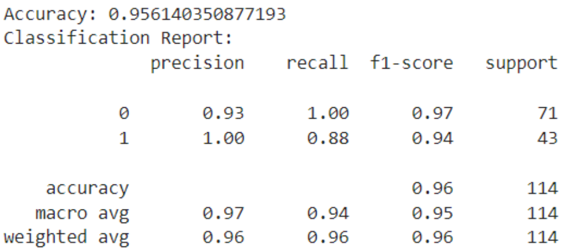


Fig 18: ROC_AUC Curve for KNN.



```
Accuracy: 0.956140350877193
Classification Report:
              precision    recall  f1-score   support

           0       0.93      1.00      0.97        71
           1       1.00      0.88      0.94        43

    accuracy                           0.96       114
   macro avg       0.97      0.94      0.95       114
weighted avg       0.96      0.96      0.96       114
```

Fig: Classification Report for KNN.

In the following table we compare all our models in terms of precision, recall ,f-1 score and accuracy. From the table we get the idea about how different model worked on our dataset.

Table 1: Comparison of Different Algorithms.

Algorithm	Class	Precision	Recall	F1-score	Accuracy
Random Forest	0	0.96	0.99	.97	96.49%
	1	1.00	0.97	0.98	
KNN	0	0.93	1.00	0.97	95.61%
	1	1.00	0.88	0.94	
Gradient Boosting	0	0.96	0.97	0.97	95.61%
	1	0.95	0.93	0.94	
SVM	0	0.95	1.00	0.97	96%
	1	1.00	0.90	0.95	
Naïve Bayes	**0**	**0.96**	**1.00**	**0.98**	**97.36%**
	1	**1.00**	**0.93**	**0.96**	

In our findings, we achieved highest accuracy with 97.36% . It was quite good compared to other models.

Naïve bayes work better with text classification or medical diagnosis but when it's about big datasets or there are complex relations of features it might not work as well.

V. Conclusions

In this thesis, we worked with breast cancer classification using two classes: benign and malignant. For this purpose, we employed different machine learning models to automatically classify breast cancer based on test data. Out of the five models we used, Naive Bayes outperformed all others, achieving an overall accuracy of 97.36%. The other four models also classified with more than 95% accuracy.

Naive Bayes demonstrated exceptional performance, particularly in its precision and recall for the benign class (0.96 and 1.00, respectively) and the malignant class (1.00 and 0.90, respectively). Its simplicity and efficiency make it an excellent choice for this dataset. However, it is important to note that Naive Bayes may not perform as well with larger and more complex datasets due to its assumptions of feature independence and normally distributed data. Future work should explore ways to mitigate these limitations and enhance its scalability and robustness.

VI. Future Work

In our future works we focus on

- Working with Deep Learning models on our dataset.
- Improve the accuracy of classification
- Working with bigger and sophisticated dataset.

References

1. Sharma, G.N., Dave, R., Sanadya, J., Sharma, P. and Sharma, K., 2010. Various types and management of breast cancer: an overview. Journal of advanced pharmaceutical technology & research, 1(2), pp.109-126.

2. World Health Organization. Breast cancer. [WHO website]. World Health Organization. Accessed [date you accessed WHO website].

3. Chauhan, A. (2021, February 24). Fully Explained Gradient Boosting Technique in Supervised Learning. Towards AI. Retrieved from https://pub.towardsai.net/fully-explained-gradient-boosting-technique-in-supervised-learning-d3e293ca70e1.

4. Soni, S. (2020, October 22). K-Nearest Neighbours: Introduction to Machine Learning Algorithms. Medium. Retrieved from https://medium.com/@sachinsoni600517/k-nearest-neighbours-introduction-to-machine-learning-algorithms-9dbc9d9fb3b2.

5. Lynch, C. (2019, August 8). Demystifying the Random Forest. Towards Data Science. Retrieved from https://towardsdatascience.com/demystifying-the-random-forest-8a46f4fd416f

6. "Support Vector Machines." Scikit-learn: Machine Learning in Python, scikit-learn.org/stable/modules/svm.html. Accessed 28 May 2024.

7. Osareh, A. and Shadgar, B., 2010, April. Machine learning techniques to diagnose breast cancer. In *2010 5th international symposium on health informatics and bioinformatics* (pp. 114-120). IEEE.

8. Yue, W., Wang, Z., Chen, H., Payne, A. and Liu, X., 2018. Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, *2*(2), p.13.

9. Fatima, N., Liu, L., Hong, S. and Ahmed, H., 2020. Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, *8*, pp.150360-150376.

10. Amrane, M., Oukid, S., Gagaoua, I. and Ensari, T., 2018, April. Breast cancer classification using machine learning. In *2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT)* (pp. 1-4). IEEE.

APPENDIX A: LITERATURE REVIEW

Introduction to use of ML in Health sector

- One of the most common and dangerous diseases impacting women globally is breast cancer. Breast cancer death rates can be significantly reduced and patient outcomes improved by early identification and precise prediction. There is increasing interest in using machine learning approaches to create prediction models for breast cancer diagnosis as a result of technological improvements and the accessibility of large-scale datasets. The manual detection of breast cancer from a huge number of variables and factors can be very time consuming and challenging. Recent studies regarding the use of ML and AI in health sector has provided significant insight into the matter where critical use of data sorting, pre-processing and various classification tools have been utilized. The shortcomings of these researches are mostly not having large datasets and unable to find a balanced dataset with positive and negative classes.
- Breast Cancer detection using ML has been implemented based on different Machine Learning classifier methodologies which include-
  - Use of SVM
  - Use of Logistic Regression
  - Use of Random Forest
  - Use of Naïve Bayes

Body of Literature Review

SVM (Support Vector Machines)

The initial introduction of Support Vector Machines (SVMs) can be attributed to Vladimir Vapnik, who proposed their use in the context of two-class categorization. This technique aims to identify the decision hyperplane that optimally maximises the separation margin between data points belonging to different classes, as described by Boser et al. (1992). The decision boundary, also known as the ideal hyperplane, is determined by the midpoint of the separation margin. The support vectors refer to the data points that are closest to this decision boundary. Support Vector Machines (SVMs) are classified within the broader category of kernel approaches. Kernel methods have the ability to function effectively in high-dimensional spaces due to their reliance on dot-products as the sole means of incorporating data.

Researchers has conducted research[1] on the classification of Breast Cancer cases using the database from “The UCI machine learning repository where there were 569 occurrences and 32 attributes where their SVM accuracy was found to be 96.49%, sensitivity 97%, specificity 95%, precision 93%, Recall 97% and F1 score of 95%. But this was not the best accuracy in their research. Instead, the highest accuracy was achieved from logistic regression. In terms of using SVM (support vector machine), it has been proven to be highly accurate in many cases such as- the research by [2] where the accuracy was 96.25%. and by [3] who found an accuracy of 96.92% paired with RBF kernel. However, researchers [4] found the use of SVM in their non-standardised dataset to be surprisingly ineffective. In that scenario, Gaussian NB, KNN and CART performed the best. But upon standardisation, the SVM performed approximately around 98% accuracy. It is to be noted that, in their research, the number of class variable was binary that is, only two classes. For multiple classes, the researcher concluded that modified version of SVM should be implemented. They also suggested that it can be implemented on a cloud platform for easier usage. Researchers [5] included two dataset in their research in predicting breast cancer cells. Upon analyzation their final dataset had 30 attributes out of 596 attributes. Evaluation for classifiers were done by 10-fold cross validation test. Upon discussing their result it was found that SVM is comparatively better than other classifiers such as KNN. SVM required 0.08 second to build the model whereas KNN required 0 second. That’s because it doesn’t function much during the training process of the dataset. In the end, SVM outperformed every other classifiers in different factors such as accuracy, sensitivity specificity etc by obtaining 97.9% accuracy. It correctly predicted 569 instances out of 699 instances with 12 incorrect predictions. Tests were run on a similar sized dataset by [6] where they had 699 cancer records with two classes of data.

Performance metrics such as True Positive(TP) False Positive(FP), True Negative (TN), False Negative(FN) were used to predict the accuracy. SVM excelled in their research in terms of accuracy which was about 97%. However, in the research [7] it was found that SVC did not provide the highest accuracy within the same sized dataset containing 569 instances. In [8] the researchers also generated actual and predicted result similar to (Ahmed et al, 2020) using their same dataset. Interestingly, their research showed SVM accuracy to be 96.5% which is lower than many research outcomes.

Logistic Regression

Logistic regression is a statistical technique commonly employed in the field of binary classification to probabilistically model occurrences or classes. Logistic regression is a statistical model commonly employed for the purpose of modelling binary classification problems. It utilizes the logistic function, and other more intricate extensions have been developed for logistic regression. Logistic regression is a supervised classification algorithm. Logistic regression is a regression model that employs regression techniques to estimate the probability of a given data object or entry belonging to a specific category [9]

Authors has done research [10] on three different breast cancer datasets using a modified version of Logistic regression ML model. A weighted sigmoid function was incorporated in the classical logistic regression model which led to drastic improvement in the overall performance of the ML model. It is to be noted that this approach proved to be effective for two different sized dataset having same features. However, the exact relationship between the sigmoid function and the data size is yet to be determined. If done correctly, it will significantly reduce the search complexity of the algorithm. Logistic regression was used on the WDBC dataset by [11]. Selecting 17 attributes from the dataset, it was found that Logistic Regression shows an accuracy of 0.97 and a precision score of 0.95. However, this is less compared to the other ML approaches in their research. In similar research on the same dataset, authors [12] found that binary regression model could prove to be very effective. The correlation matrix was used to determine the highly correlated features of the dataset. Next, the regression model was built from an empty model and the least correlated features were added. This method showed an average classification accuracy score of 98.9%, 98.5% sensitivity score and 99.1% specificity score. Using three datasets including WBCD, WDBC and WPBC, authors [13] implemented logistic regression for the feature selection process of the dataset. Although the classification was done using GMDH classifier, using logistic regression in the process of feature selection proved to be quite significant. GMDH outperformed most other previous works on the same dataset. The accuracy of this approach was 97.9%, 99.1% and 84.6% in the WBCD, WDBC and WPBC datasets respectively.

Future research should explore the relationship between the weighted sigmoid function and data size, as its effectiveness in logistic regression remains unclear. Additionally, comprehensive comparisons with other ML models and optimization for large-scale data are needed. Investigating various feature selection techniques, hybrid approaches with other algorithms, and strategies for handling imbalanced data could provide valuable insights. The impact of data preprocessing on performance and maintaining model interpretability, especially with complex modifications, are also crucial areas for study. Addressing these gaps can enhance the understanding and application of logistic regression across different domains.

Random Forest

In the Random Forest algorithm, a substantial quantity of decision trees is generated. Each observation is inputted into every decision tree. The final output is determined by selecting the most frequently occurring outcome for each observation. The newly acquired observation is inputted into all of the trees, and a majority vote is conducted for each categorization model. An error calculation is conducted for the classes that were not utilized in the construction of the tree. This phenomenon is commonly referred to as an out-of-bag error estimate, which is typically expressed as a percentage [14]

The researchers [14] compared different ML algorithms including Linear Regression, Decision Tree and Random Forest. They created an error estimate for the case where trees were not built. This is called the OOB (Out-of-bag) error estimate. In their case study, regression tree was chose for random forest analysis. The number of trees that were considered numbered about 500 and the variables were split into 2. The total percentage of the variance of the attributes that were considered in the random forest was 88.14%. This is called the residual sum of squares (RSS). Simply put, the prediction percentage for the random forest was 88.14%. For future work, more pre-processing needs to be done on the data. AdaBoost, sometimes known as Adaptive Boost, is an abbreviation used to refer to the algorithm. In the context of classification issues, the algorithm transforms a collection of weak classifiers into a single, robust classifier. The classifier should undergo iterative training on a range of weighted training examples. With each cycle, the classifier consistently produces valuable outcomes by effectively reducing the training error. Authors [15] implemented AdaBoost in their research on breast cancer detection using Random Forest. AdaBoost, sometimes known as Adaptive Boost, is an abbreviation used to refer to the algorithm. In the context of classification issues, the algorithm transforms a collection of weak classifiers into a single, robust classifier. The classifier should undergo iterative training on a range of weighted training examples. With each cycle, the classifier consistently produces valuable outcomes by effectively reducing the training error. Adaboost algorithm improves upon the performance in binary classification. If it is combined with 10 fold cross validation, it performs even better in terms of accuracy. The proposed method shows an accuracy of 98.5714% whereas in [16], the authors found the highest testing accuracy to be 97.1429% using LPSVM. However, in [17], the Sensitivity, Specificity and Accuracy for Random Forest classifier was found to be 75%, 70% and 72% respectively in 250 testing cases. The research conducted by [18] presents a novel approach aimed at improving the classification performance of breast cancer datasets. The authors employed an innovative methodology that utilized Kernel Neutrosophic C-Means Clustering (KNCMC) to assign weights to features and optimized the Random Decision Forest (RDF) classifier model by the application of the Bayesian Optimization algorithm. The FW + BOA-RDF approach demonstrated significantly improved accuracy when used to the Breast Cancer Wisconsin (Prognosis) Data Set. In the evaluation conducted on the WPBC dataset, using a training set including 75% of the data and a testing set comprising 25% of the data, the FW + BOA-RDF technique demonstrated superior performance compared to other current methods such as FW + ALO-BPNN, FW + SSA-SVM, and FW + GA-SVM. The margins of improvement achieved by the FW + BOA-RDF method over these methods were 3.7146%, 5.27398%, and 4.4413% respectively.

Future research on Random Forest should focus on enhanced data preprocessing and exploring hybrid models like AdaBoost combined with cross-validation to improve accuracy. Additionally, investigating methods to optimize feature weighting and classifier performance, such as Kernel Neutrosophic C-Means Clustering (KNCMC) with Bayesian Optimization, could yield better results. Comparative analyses with other ML algorithms and strategies to handle varying dataset characteristics and size are crucial. Understanding and minimizing out-of-bag errors, improving variance explanations, and achieving higher prediction accuracy are key areas for development. Addressing these gaps can significantly advance the application of Random Forest in complex classification tasks like breast cancer detection.

Naïve Bayes

The Naive Bayes algorithm is widely recognized as a highly efficient method for statistical and probabilistic classification. The health care environment is characterized by an abundance of information but a lack of expertise. In order to acquire knowledge, it is imperative to employ efficient analytical techniques that are designed to unveil concealed links inside datasets [19]

Another research was conducted by [29] on breast cancer dataset. This study aimed to identify the most minimal set of indicators that can guarantee a high level of accuracy in classifying breast cancer as either benign or malignant. A comparison analysis was undertaken to evaluate various cancer classification methods, including Naïve Bayes, Support Vector Machine, and Ensemble classifiers. The study also examined the time complexity of each classifier. The Naïve Bayes classifier was determined to be the most effective classifier because to its significantly lower temporal complexity in comparison to the other two classifiers. The authors used binning technique on the pre-processed data where all the range of values of each attribute is divided into three bins. The project showed the Naïve Bayes accuracy of 97.3978% with five dominant features and time complexity of 0.102023 millisecond with was least compared to the other classifiers. The primary aim of the study conducted by [19] was to create a breast cancer prediction system based on Naive Bayes Classifiers. The system was intended to assist experts in making accurate decisions. The technology has the potential to be deployed in remote areas such as rural regions or countryside locations, with the aim of replicating human diagnostic skill in the treatment of cancer. Due to an existing model, the system is user-friendly and dependable. The training process involved the utilization of Wisconsin Datasets, which consisted of 699 entries comprising 9 medical attributes. A total of 200 recordings were collected for the purpose of conducting tests. The dataset consisted of approximately 65.5% benign cases and the remaining 34.5% were classified as malignant cases. The study determined that the level of accuracy achieved was 93%. In another research by [20], comparisons were made between Naïve Bayes, RBF Network and Decision Tree algorithm on the Wisconsin dataset. The dataset used in this study was obtained from the UCI Machine Learning repository and consisted of 683 occurrences. Data selection, preprocessing, and transformation techniques were employed in order to construct the prediction models. Here, a binary categorical variable was used to represent the survival outcome of the variable where malignant was denoted as "1" and benign was marked as "0". To measure the unbiased prediction accuracy for all three models, a 10 fold cross validation method was used. This approach showed a classification accuracy of 97.36% for Naïve Bayes classifier which was the highest. Similarly in [21] Gaussian Naïve Bayes was implemented to detect both Breast cancer and Lung cancer from the Wisconsin Breast Cancer Dataset. The research showed that the method to be highly effective as the accuracy obtained from Gaussian Naïve Bayes was 98%, Sensitivity was 97% and specificity was 98%. In this study [22], the authors examined the efficacy of several supervised learning algorithms, including Naive Bayes, Support Vector Machine (SVM) with Gaussian RBF kernel, Radial

Basis Function (RBF) neural networks, Decision tree J48, and basic CART. The algorithms discussed in this study were utilized for the purpose of classifying breast cancer datasets, specifically WBC, WDBC, and Breast tissue datasets obtained from the UCI Machine Learning Repository . The trials were conducted utilizing the WEKA tool. The Naive Bayes algorithm demonstrates an accuracy rate of 96.50% for the WBC dataset, 94.33% for the Breast tissue dataset, and 92.61% for the WDBC dataset.

Future research on Naive Bayes should address its limitations with large datasets and explore ways to improve its performance. While it shows high efficiency and accuracy in breast cancer classification, particularly with minimal feature sets and low time complexity, its assumptions of feature independence and normally distributed data can be problematic. Studies should focus on hybrid approaches and advanced preprocessing techniques to mitigate these issues. Comparative analyses with other classifiers like SVM and Decision Trees are necessary to understand its relative strengths and weaknesses. Additionally, enhancing its applicability in remote healthcare settings can further its practical utility. Addressing these gaps can optimize Naive Bayes for broader and more complex datasets, improving its reliability in diverse scenarios.

Other

A study titled "Breast Cancer Diagnosis Utilizing an Adaptive Voting Ensemble Machine Learning Algorithm" [23] was conducted. The utilization of ensemble methods is employed in the diagnosis of breast cancer through the integration of neural networks and logistic algorithms. Ensemble approaches leverage the utilization of several models in order to get enhanced outcomes. Ensemble models typically yield more precise outcomes compared to individual models. The most straightforward ensemble models are the 'Voting model' and the 'Averaging model' because to their simplicity in comprehension and implementation. The averaging model is commonly employed in regression tasks, while the Voting model is typically utilized in classification tasks. The data was pre-processed using the Standardization method, and the features were chosen based on the 'Univariate feature selection' approach, which identifies the most optimal features through univariate statistical tests. The process involves assessing the relationship between each feature and the target variable in order to determine the predictive capability of the given feature with respect to the target variable. The term used to refer to this statistical technique is analysis of variance (ANOVA). This nomenclature is attributed to the term 'univariate'. Every each characteristic possesses its own corresponding evaluation score. Lastly, the Neural Network model was employed for the purpose of cancer diagnosis. The aforementioned strategies utilizing neural networks offer several advantages, including enhanced prediction accuracies. However, it is important to note that these techniques also have a drawback in that they require a longer processing time due to the inherent complexity of neural network algorithms. In a separate scholarly publication titled "Breast Cancer Diagnosis Utilizing Deep Learning Algorithm" [24] deep learning algorithms were employed for the purpose of diagnosing breast cancer. The convolutional neural network was employed for the purpose of cancer diagnosis. Deep learning is a specialized branch of machine learning that use neural network methods inspired by the human brain to acquire knowledge from extensive datasets, hence facilitating the resolution of intricate issues. The researchers employed data preprocessing approaches, including the utilization of the Label Encoder Method. This method facilitates the conversion of non-numeric labels into a numeric format, hence enabling their utilization in machine learning models. In order to enhance the accuracy of label assignment, machine learning methods can be employed. The preprocessing phase discussed here holds significant importance in the context of conducting supervised learning. The normalization and standard scalar method is based on the assumption that the data follows a normal distribution within each feature. This method scales the data in such a way that the distribution is centred around 0 and has a standard deviation of 1. When the distribution of data is not normal, it is advisable to avoid using certain scalars as they have the potential to impact the outcomes. Subsequently, a deep learning neural network method was employed, comprising a sequence of algorithms designed to identify meaningful correlations within the dataset, emulating the cognitive processes of the human brain, with the objective of cancer diagnosis.

The paper in reference [25] presents a detailed analysis of an automated approach for identifying irregularities in mammograms. The segmentation of the suspect region-of-interest (ROI) was performed using the fuzzy-C-means algorithm and a thresholding approach. The algorithm developed for the Mini-MIAS dataset underwent validation. The researchers reached the conclusion that the detection of worrisome regions in mammograms can be enhanced by removing pre-processed enhanced and pre-processed enhanced inverted pictures. In their study, the authors presented a method, as described in reference [26], for the purpose of distinguishing between malignant and benign states. This algorithm relies on a fuzzy inference system. The comparison of traditional performance metrics, such as sensitivity, accuracy, and specificity, indicates that the proposed solution demonstrates superior performance compared to the classification methods of Artificial Neural Network (ANN) and Support Vector Machine (SVM). The study conducted by Katsis et al. [27] employed a methodology that involved utilizing a Correlation

Feature Selection (CFS) procedure to prioritize various retrieved characteristics. Additionally, an Artificial Immune Recognition System (AIRS) classifier was employed to aid in the identification of breast cancer. In order to assess the methodology, data was collected from a sample of 53 participants out of a total of 4726 cases. The specific subjects discussed in the text are lesions that did not exhibit strong indications of being either benign or malignant when assessed using all available modalities. In each instance, a biopsy was performed and the biopsy findings were utilized as the benchmark to authenticate the process. The dataset that was created included both the characteristics and the biopsy results (indicating whether the person had malignancy or benignity) for all 53 individuals.

All data were collected at the University Hospital of Ioannina, Greece. The Support Vector Machine (SVM) technique yielded an accuracy of 70.00+6.33% when applied to the complete collection of features. Additionally, when examining only the subset of features selected using the Correlation-based Feature Selection (CFS) method, the accuracy achieved was 68.92+6.97%. In the study conducted by [28], various predictive models including Logistic regression, random forests, Support vector machines, Artificial neural networks, and ensembles were employed for the purpose of diagnosing breast cancer. The performance measurements employed in this study encompass accuracy, the area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, and specificity. Subsequently, four predictive models are refined and subsequently integrated through ensemble approaches in order to attain an enhanced predictive model. The accuracy of the ensemble model in the final evaluation is 98.23%.

Multiple research projects have employed the WBC dataset as a means of detecting breast cancer. The research conducted by Karabatak et al. [30] focuses on the detection of breast cancer using a combination of association rules and a neural network. Association rules are employed to eliminate superfluous data, hence lowering the dimensionality of features. The neural network employs the remaining features to classify each report. The ultimate model yielded a precision rate of 97.4%.

Future research on breast cancer diagnosis should focus on improving ensemble methods, preprocessing techniques, and the integration of advanced algorithms. Studies have shown that ensemble models like Voting and Averaging yield higher accuracy than individual models. Exploring hybrid approaches with neural networks and logistic regression, and optimizing preprocessing methods like Standardization and Univariate feature selection, can enhance performance. Additionally, deep learning models such as convolutional neural networks (CNNs) have shown promise but require extensive data preprocessing and longer processing times. Addressing these limitations and exploring novel algorithms like fuzzy inference systems and Artificial Immune Recognition Systems can further improve accuracy and robustness in breast cancer detection.

Conclusions from the Literature Review

To conclude that upon analysis of the aforementioned literature reviews, there are still a lot of scopes for better understanding of the attributes related to breast cancer and data pre-processing. Also, based on the size of database, the ML algorithms will vary their accuracy. So various trial and error has to be conducted on balancing the data using pre-processing methods for better output of the ML classifiers. More research papers need to be examined to understand the gaps in the research of this particular field in health informatics.

References of the Literature Review

1. Jamal, J. H. Antor, R. Kumar and P. Rani, "Breast Cancer Prediction Using Machine Learning Classifiers," 2022 5th International Conference on Advances in Science and Technology (ICAST), Mumbai, India, 2022, pp. 456-459, doi: 10.1109/ICAST55766.2022.10039656.
2. V. A. Telsang and K. Hegde, "Breast Cancer Prediction Analysis using Machine Learning Algorithms," 2020 International Conference on Communication, Computing and Industry 4.0 (C2I4), Bangalore, India, 2020, pp. 1-5, doi: 10.1109/C2I451079.2020.9368911.
3. A. Mangal and V. Jain, "Prediction of Breast Cancer using Machine Learning Algorithms," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2021, pp. 464-466, doi: 10.1109/I-SMAC52330.2021.9640813.
4. A. Bharat, N. Pooja and R. A. Reddy, "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis," 2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C), Bangalore, India, 2018, pp. 1-4, doi: 10.1109/CIMCA.2018.8739696.
5. Y. Khourdifi and M. Bahaj, "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification," 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Kenitra, Morocco, 2018, pp. 1-5, doi: 10.1109/ICECOCS.2018.8610632.
6. M. R. Ahmed, M. A. Ali, J. Roy, S. Ahmed and N. Ahmed, "Breast Cancer Risk Prediction based on Six Machine Learning Algorithms," 2020 IEEE Asia-Pacific Conference on

Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 2020, pp. 1-5, doi: 10.1109/CSDE50874.2020.9411572.

7. S. K. Mohapatra, A. Jain, Anshika and P. Sahu, "Comparative Approaches by using Machine Learning Algorithms in Breast Cancer Prediction," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 1874-1878, doi: 10.1109/ICACITE53722.2022.9823470.
8. S. Ara, A. Das and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," 2021 International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 2021, pp. 97-101, doi: 10.1109/ICAI52203.2021.9445249.
9. H. Yusuff, N. Mohamad, U.K. Ngah and A.S. Yahaya, "Breast Cancer Analysis using Logistic Regression", International Journal of Recent Research and Applied Studies, vol. 10, no. 1, pp. 14-22, Jan 2012.
10. Laila Khairunnahar, Mohammad Abdul Hasib, Razib Hasan Bin Rezanur, Mohammad Rakibul Islam, Md Kamal Hosain, Classification of malignant and benign tissue with logistic regression, Informatics in Medicine Unlocked, Volume 16, 2019, 100189, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2019.100189.
11. R. MurtiRawat, S. Panchal, V. K. Singh and Y. Panchal, "Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 534-540, doi: 10.1109/ICESC48915.2020.9155783.
12. F. Seddik and D. M. Shawky, "Logistic regression model for breast cancer automatic diagnosis," 2015 SAI Intelligent Systems Conference (IntelliSys), London, UK, 2015, pp. 150-154, doi: 10.1109/IntelliSys.2015.7361138.
13. Ziba Khandezamin, Marjan Naderan, Mohammad Javad Rashti, Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier, Journal of Biomedical Informatics, Volume 111, 2020, 103591, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2020.103591. (https://www.sciencedirect.com/science/article/pii/S1532046420302173)
14. S. Murugan, B. M. Kumar and S. Amudha, "Classification and Prediction of Breast Cancer using Linear Regression, Decision Tree and Random Forest," 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Mysore, India, 2017, pp. 763-766, doi: 10.1109/CTCEEC.2017.8455058.

15. T. I. Rohan, Awan-Ur-Rahman, A. B. Siddik, M. Islam and M. S. U. Yusuf, "A Precise Breast Cancer Detection Approach Using Ensemble of Random Forest with AdaBoost," 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 2019, pp. 1-4, doi: 10.1109/IC4ME247184.2019.9036697.
16. Azar, Ahmad Taher & Elsaid, Shaimaa. (2012). Probabilistic neural network for breast cancer classification. Neural Computing and Applications. 23. 1737-1751. 10.1007/s00521-012-1134-8.
17. F. K. Ahmad and N. Yusoff, "Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier," 2013 13th International Conference on Intellient Systems Design and Applications, Salangor, Malaysia, 2013, pp. 121-125, doi: 10.1109/ISDA.2013.6920720.
18. Pratheep Kumar P, Mary Amala Bai V, Geetha G. Nair, An efficient classification framework for breast cancer using hyper parameter tuned Random Decision Forest Classifier and Bayesian Optimization, Biomedical Signal Processing and Control, Volume 68, 2021, 102682, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2021.102682. (https://www.sciencedirect.com/science/article/pii/S1746809421002792)
19. Kharya, S., Agrawal, S., & Soni, S. (2014). Naive Bayes classifiers: a probabilistic detection model for breast cancer. Int. J. Comput. Appl, 92(10), 26-31.
20. Chaurasia V, Pal S, Tiwari B. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology. 2018;12(2):119-126. doi:10.1177/1748301818756225
21. H. Kamel, D. Abdulah and J. M. Al-Tuwaijari, "Cancer Classification Using Gaussian Naive Bayes Algorithm," 2019 International Engineering Conference (IEC), Erbil, Iraq, 2019, pp. 165-170, doi: 10.1109/IEC47844.2019.8950650.
22. Aruna, S., Rajagopalan, S. P., & Nandakishore, L. V. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer. Computer Science & Information Technology, 2(2011), 37-45.
23. Naresh Khuriwal and Nidhi Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm", *IEEMA Engineer Infinite Conference (eTechNxT)*, 2018.

24. Naresh Khuriwal and Nidhi Mishra, "Breast Cancer Diagnosis Using Deep Learning Algorithm", *2018 International Conference on Advances in Computing Communication Control and Networking (ICACCCN)*, 2018
25. K. L. Kashyap, M. K. Bajpai, and P. Khanna, "Breast cancer detection in digital mammograms," in 2015 IEEE international conference on imaging systems and techniques (IST). IEEE, 2015
26. F.-T. Johra and M. M. H. Shuvo, "Detection of breast cancer from histopathology image and classifying benign and malignant state using fuzzy logic," in 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT). IEEE, 2016, pp. 1–5

APPENDIX B: DATA AND PROGRAMS USED FOR THIS REPORT[1]

Data Preprocessing:

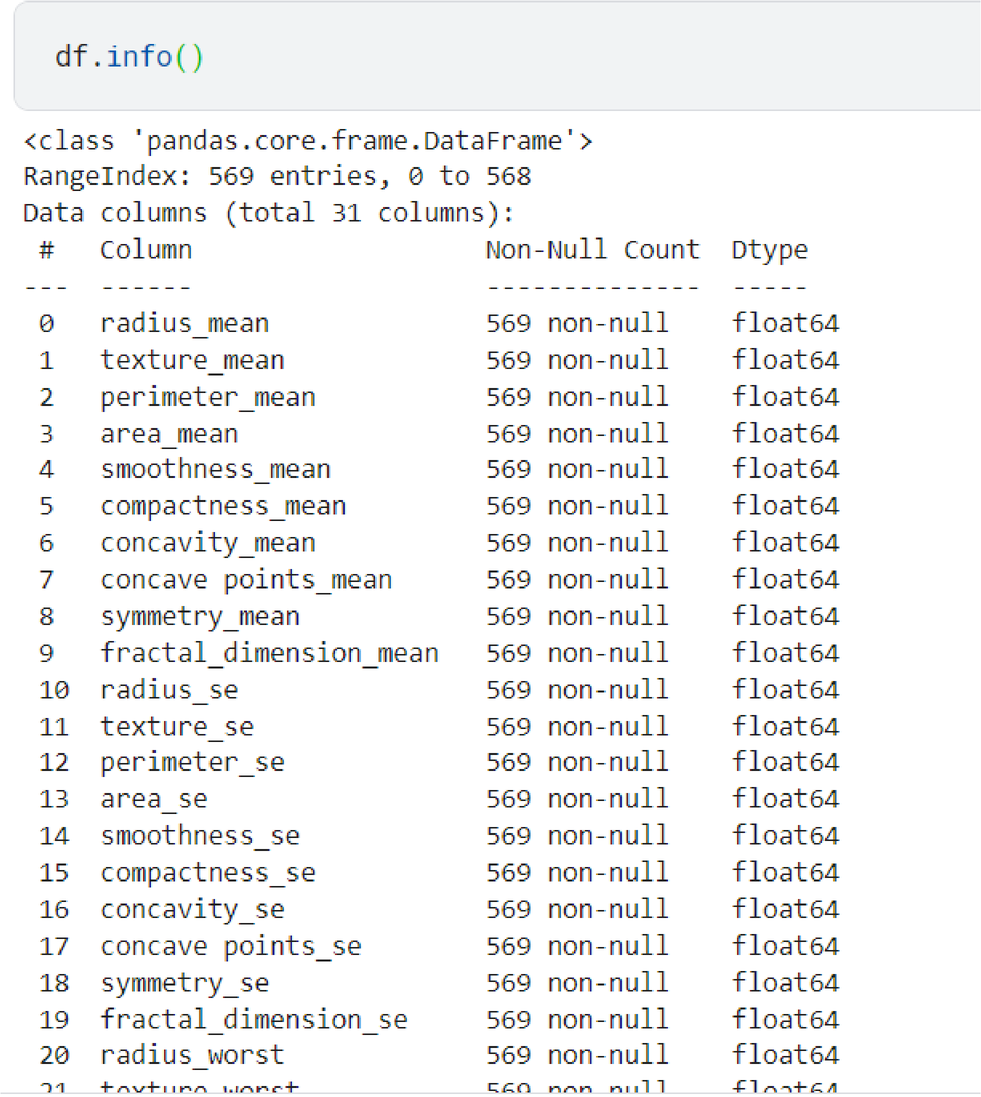
```
import pandas as pd

# Load the dataset into a DataFrame
df = pd.read_csv('/kaggle/input/breast-cancer/breastcancerdata.csv')
from sklearn.preprocessing import LabelEncoder
# Initialize the label encoder
label_encoder = LabelEncoder()

# Fit and transform the 'diagnosis' column
df['diagnosis_encoded'] = label_encoder.fit_transform(df['diagnosis'])
# Drop the 'diagnosis' column
df.drop(columns=['diagnosis'], inplace=True)

# Rename the 'diagnosis_encoded' column to 'diagnosis'
df.rename(columns={'diagnosis_encoded': 'diagnosis'}, inplace=True)


X = df.drop(columns=['diagnosis'])  # Features
y = df['diagnosis']  # Target variable
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report
# Split the data into training and testing sets

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
```

Model Evaluation:

```
# Make predictions on the testing data
y_pred = rf_classifier.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
print("Classification Report:")
print(classification_report(y_test, y_pred))
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay,
roc_curve, roc_auc_score

# Confusion Matrix
cm = confusion_matrix(y_test, y_pred)
cmd = ConfusionMatrixDisplay(cm, display_labels=rf_classifier.classes_)
cmd.plot()
plt.title('Confusion Matrix')
plt.show()

# ROC-AUC Curve
```

```
y_prob = rf_classifier.predict_proba(X_test)[:, 1]  # Get the probability of the positive class
fpr, tpr, thresholds = roc_curve(y_test, y_prob)
roc_auc = roc_auc_score(y_test, y_prob)

plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], color='skyblue', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc="lower right")
plt.show()

# Display the results
print("Accuracy:", accuracy)
print("Classification Report:")
print(classification_report(y_test, y_pred))
```

APPENDIX C: FIGURES[2]



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 31 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   radius_mean              569 non-null    float64
 1   texture_mean             569 non-null    float64
 2   perimeter_mean           569 non-null    float64
 3   area_mean                569 non-null    float64
 4   smoothness_mean          569 non-null    float64
 5   compactness_mean         569 non-null    float64
 6   concavity_mean           569 non-null    float64
 7   concave points_mean      569 non-null    float64
 8   symmetry_mean            569 non-null    float64
 9   fractal_dimension_mean   569 non-null    float64
 10  radius_se                569 non-null    float64
 11  texture_se               569 non-null    float64
 12  perimeter_se             569 non-null    float64
 13  area_se                  569 non-null    float64
 14  smoothness_se            569 non-null    float64
 15  compactness_se           569 non-null    float64
 16  concavity_se             569 non-null    float64
 17  concave points_se        569 non-null    float64
 18  symmetry_se              569 non-null    float64
 19  fractal_dimension_se     569 non-null    float64
 20  radius_worst             569 non-null    float64
 21  texture_worst            569 non-null    float64
```

FIG 1: DATASET INFORMATION.

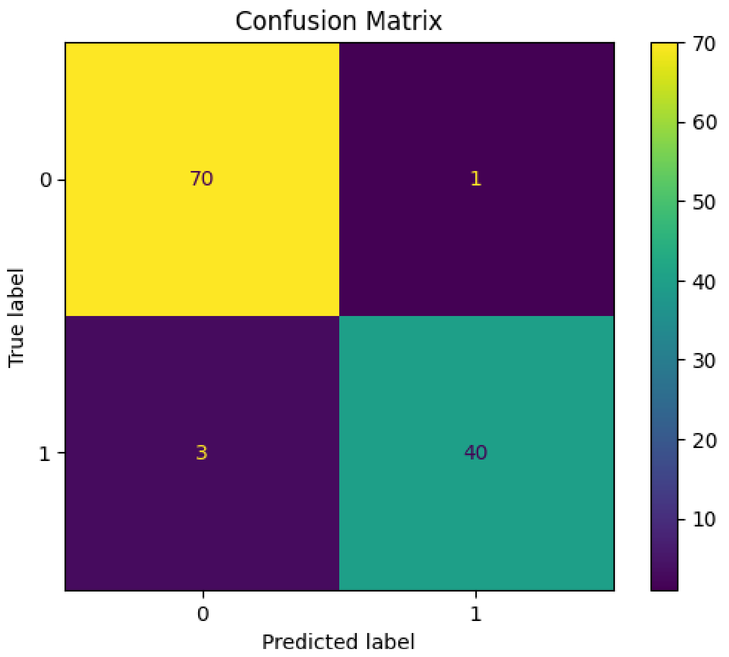


FIG 2: CONFUSION MATRIX FOR RANDOM FOREST.

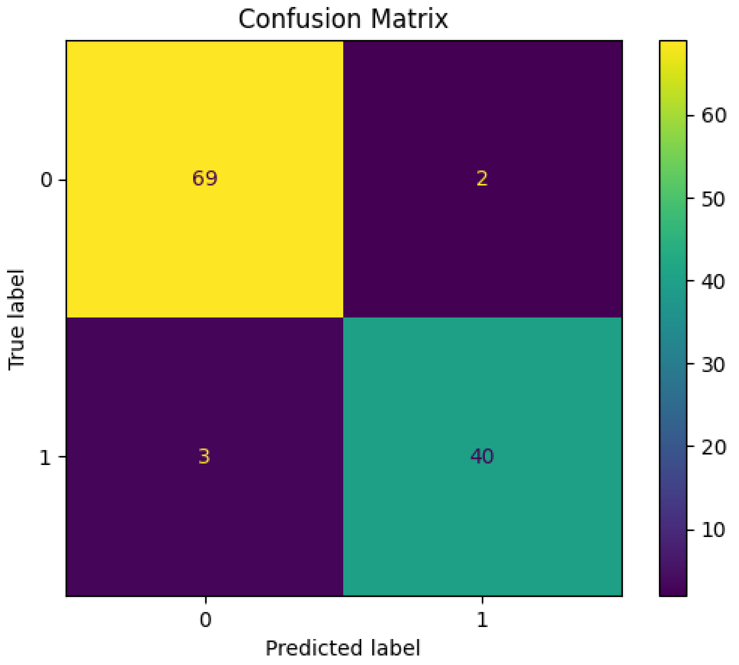


FIG 3: CONFUSION MATRIX FOR GBM.

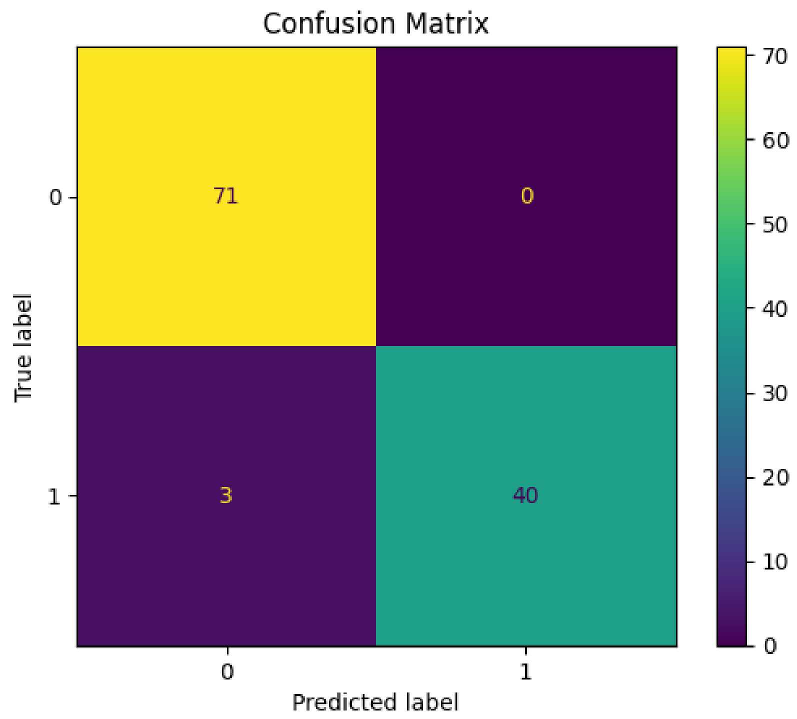


FIG 4: CONFUSION MATRIX FOR NAÏVE BAYES.

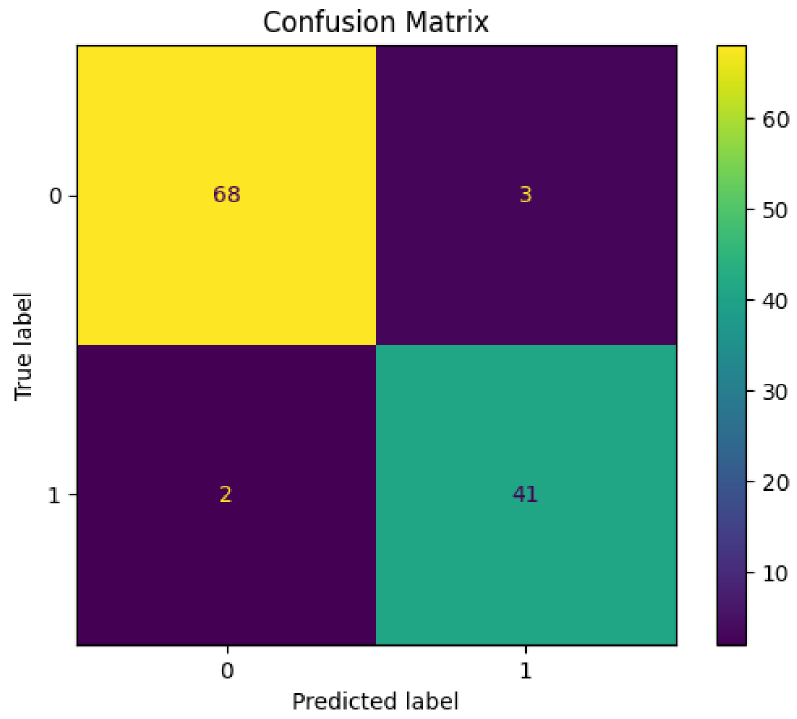


FIG 5: CONFUSION MATRIX FOR SVM.

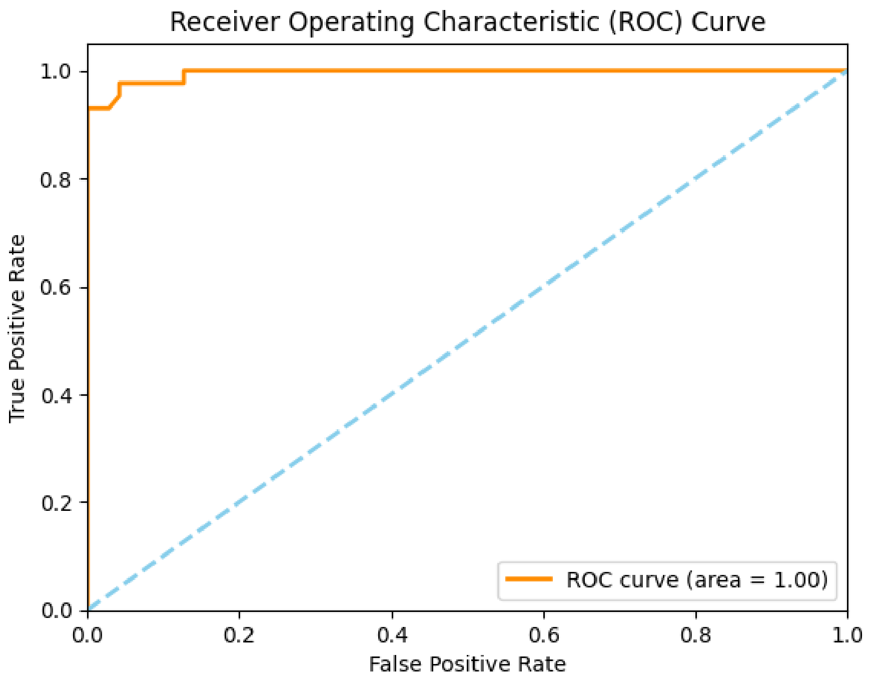


FIG6: ROC_AUC FOR GBM.

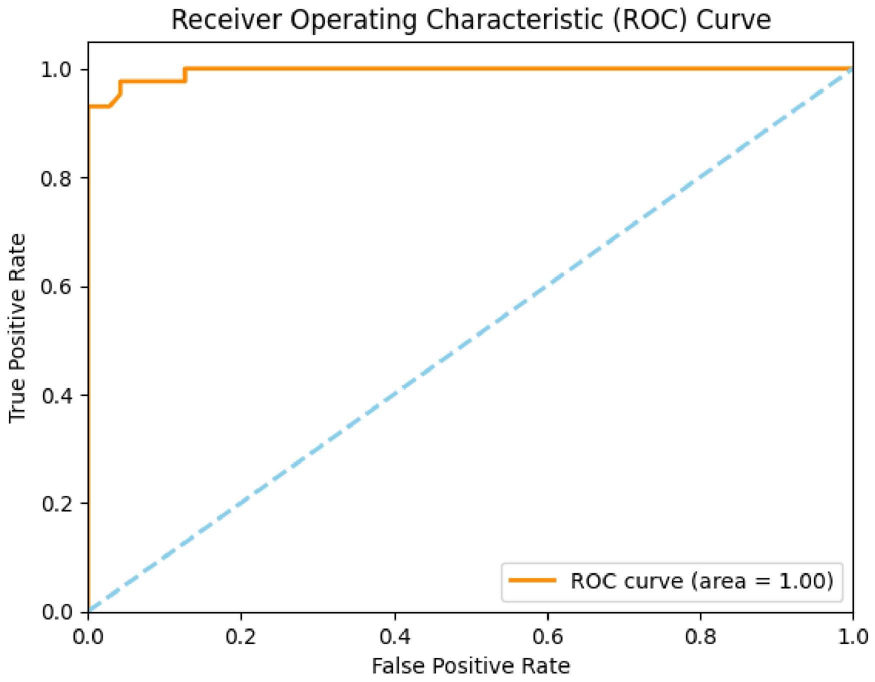


FIG6: ROC_AUC FOR KNN.

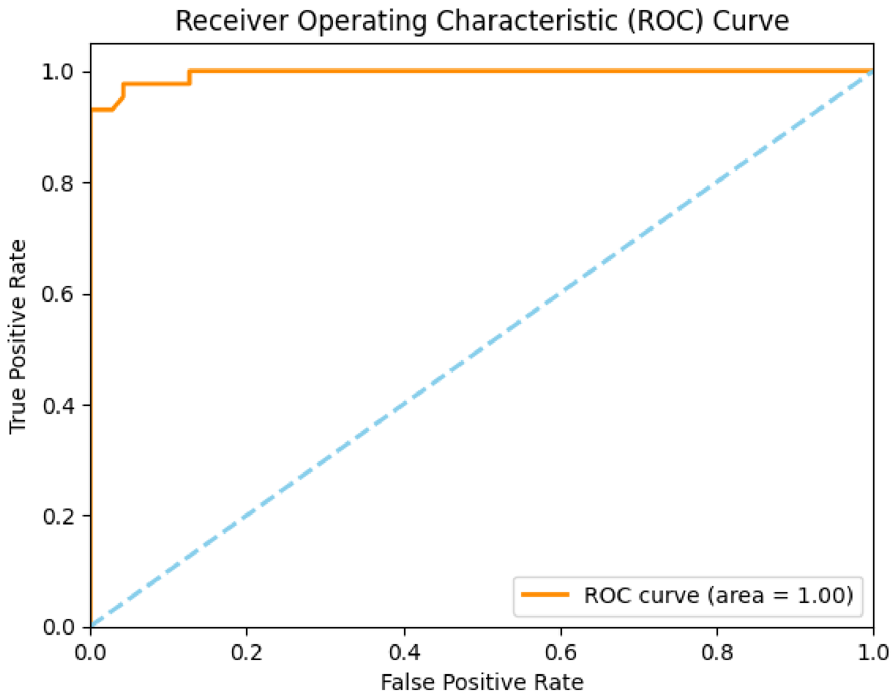


FIG8: ROC_AUC FOR NAÏVE BAYES.

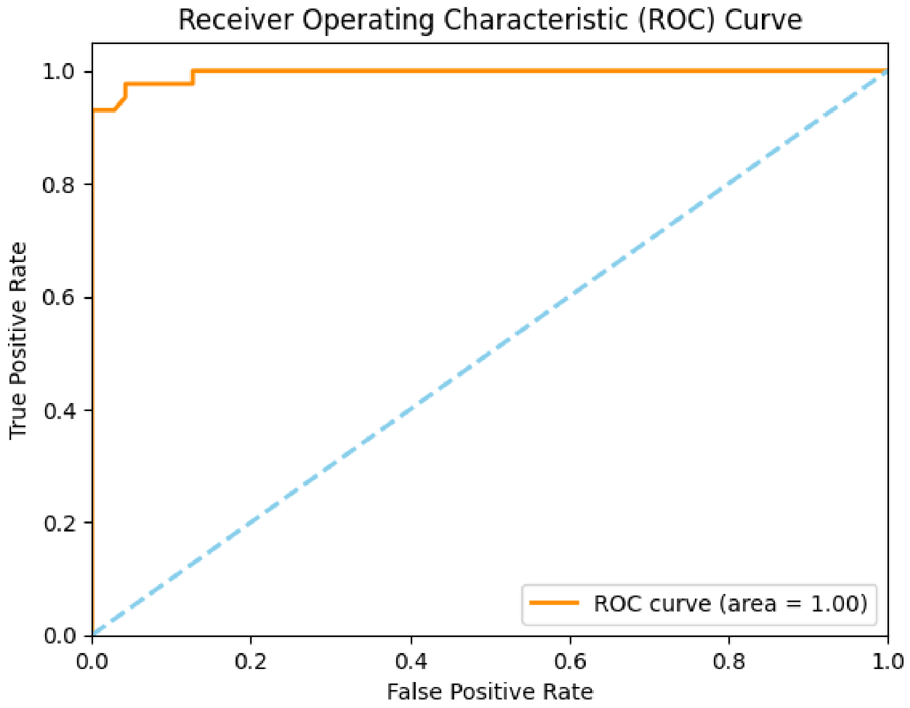


FIG9: ROC_AUC FOR RANDOM FOREST.

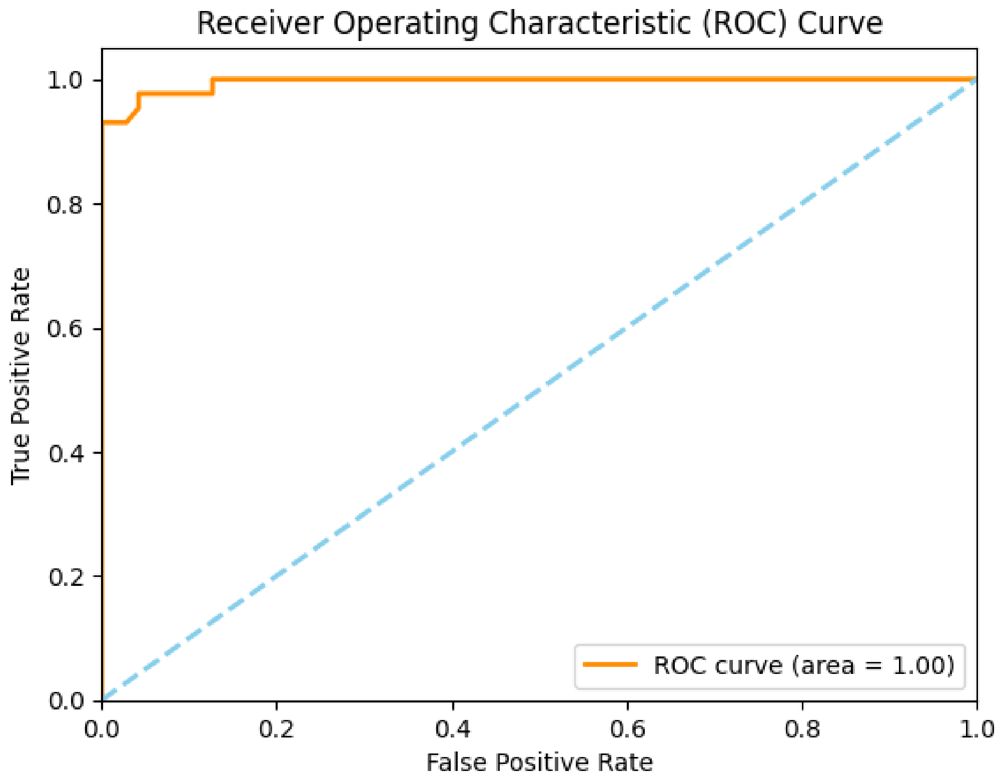


FIG: ROC_AUC FOR SVM.

APPENDIX D: TABLES[3]

Table 1: Comparison of Different Algorithms.

Algorithm	Class	Precision	Recall	F1-score	Accuracy
Random Forest	0	0.96	0.99	.97	96.49%
	1	1.00	0.97	0.98	
KNN	0	0.93	1.00	0.97	95.61%
	1	1.00	0.88	0.94	
Naïve Bayes	0	0.96	1.00	0.98	97.36%
	1	1.00	0.93	0.96	
Gradient Boosting	0	0.96	0.97	0.97	95.61%
	1	0.95	0.93	0.94	
SVM	0	0.95	1.00	0.97	96%

REFERENCES[4]

1. Jamal, J. H. Antor, R. Kumar and P. Rani, "Breast Cancer Prediction Using Machine Learning Classifiers," 2022 5th International Conference on Advances in Science and Technology (ICAST), Mumbai, India, 2022, pp. 456-459, doi: 10.1109/ICAST55766.2022.10039656.
2. V. A. Telsang and K. Hegde, "Breast Cancer Prediction Analysis using Machine Learning Algorithms," 2020 International Conference on Communication, Computing and Industry 4.0 (C2I4), Bangalore, India, 2020, pp. 1-5, doi: 10.1109/C2I451079.2020.9368911.
3. A. Mangal and V. Jain, "Prediction of Breast Cancer using Machine Learning Algorithms," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2021, pp. 464-466, doi: 10.1109/I-SMAC52330.2021.9640813.
4. A. Bharat, N. Pooja and R. A. Reddy, "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis," 2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C), Bangalore, India, 2018, pp. 1-4, doi: 10.1109/CIMCA.2018.8739696.
5. Y. Khourdifi and M. Bahaj, "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification," 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Kenitra, Morocco, 2018, pp. 1-5, doi: 10.1109/ICECOCS.2018.8610632.
6. M. R. Ahmed, M. A. Ali, J. Roy, S. Ahmed and N. Ahmed, "Breast Cancer Risk Prediction based on Six Machine Learning Algorithms," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 2020, pp. 1-5, doi: 10.1109/CSDE50874.2020.9411572.
7. S. K. Mohapatra, A. Jain, Anshika and P. Sahu, "Comparative Approaches by using Machine Learning Algorithms in Breast Cancer Prediction," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 1874-1878, doi: 10.1109/ICACITE53722.2022.9823470.
8. S. Ara, A. Das and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," 2021 International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 2021, pp. 97-101, doi: 10.1109/ICAI52203.2021.9445249.
9. H. Yusuff, N. Mohamad, U.K. Ngah and A.S. Yahaya, "Breast Cancer Analysis using Logistic Regression", International Journal of Recent Research and Applied Studies, vol. 10, no. 1, pp. 14-22, Jan 2012.
10. Laila Khairunnahar, Mohammad Abdul Hasib, Razib Hasan Bin Rezanur, Mohammad Rakibul Islam, Md Kamal Hosain, Classification of malignant and benign tissue with logistic regression, Informatics in Medicine Unlocked, Volume 16, 2019, 100189, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2019.100189.
11. R. MurtiRawat, S. Panchal, V. K. Singh and Y. Panchal, "Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 534-540, doi: 10.1109/ICESC48915.2020.9155783.

12. F. Seddik and D. M. Shawky, "Logistic regression model for breast cancer automatic diagnosis," 2015 SAI Intelligent Systems Conference (IntelliSys), London, UK, 2015, pp. 150-154, doi: 10.1109/IntelliSys.2015.7361138.
13. Ziba Khandezamin, Marjan Naderan, Mohammad Javad Rashti, Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier, Journal of Biomedical Informatics, Volume 111, 2020, 103591, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2020.103591. (https://www.sciencedirect.com/science/article/pii/S1532046420302173)
14. S. Murugan, B. M. Kumar and S. Amudha, "Classification and Prediction of Breast Cancer using Linear Regression, Decision Tree and Random Forest," 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Mysore, India, 2017, pp. 763-766, doi: 10.1109/CTCEEC.2017.8455058.
15. T. I. Rohan, Awan-Ur-Rahman, A. B. Siddik, M. Islam and M. S. U. Yusuf, "A Precise Breast Cancer Detection Approach Using Ensemble of Random Forest with AdaBoost," 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 2019, pp. 1-4, doi: 10.1109/IC4ME247184.2019.9036697.
16. Azar, Ahmad Taher & Elsaid, Shaimaa. (2012). Probabilistic neural network for breast cancer classification. Neural Computing and Applications. 23. 1737-1751. 10.1007/s00521-012-1134-8.
17. F. K. Ahmad and N. Yusoff, "Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier," 2013 13th International Conference on Intellient Systems Design and Applications, Salangor, Malaysia, 2013, pp. 121-125, doi: 10.1109/ISDA.2013.6920720.
18. Pratheep Kumar P, Mary Amala Bai V, Geetha G. Nair, An efficient classification framework for breast cancer using hyper parameter tuned Random Decision Forest Classifier and Bayesian Optimization, Biomedical Signal Processing and Control, Volume 68, 2021, 102682, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2021.102682. (https://www.sciencedirect.com/science/article/pii/S1746809421002792)
19. Kharya, S., Agrawal, S., & Soni, S. (2014). Naive Bayes classifiers: a probabilistic detection model for breast cancer. Int. J. Comput. Appl, 92(10), 26-31.
20. Chaurasia V, Pal S, Tiwari B. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology. 2018;12(2):119-126. doi:10.1177/1748301818756225
21. H. Kamel, D. Abdulah and J. M. Al-Tuwaijari, "Cancer Classification Using Gaussian Naive Bayes Algorithm," 2019 International Engineering Conference (IEC), Erbil, Iraq, 2019, pp. 165-170, doi: 10.1109/IEC47844.2019.8950650.
22. Aruna, S., Rajagopalan, S. P., & Nandakishore, L. V. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer. Computer Science & Information Technology, 2(2011), 37-45.
23. Naresh Khuriwal and Nidhi Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm", *IEEMA Engineer Infinite Conference (eTechNxT)*, 2018.
24. Naresh Khuriwal and Nidhi Mishra, "Breast Cancer Diagnosis Using Deep Learning Algorithm", *2018 International Conference on Advances in Computing Communication Control and Networking (ICACCCN)*, 2018
25. K. L. Kashyap, M. K. Bajpai, and P. Khanna, “Breast cancer detection in digital mammograms,” in 2015 IEEE international conference on imaging systems and techniques (IST). IEEE, 2015

26. F.-T. Johra and M. M. H. Shuvo, “Detection of breast cancer from histopathology image and classifying benign and malignant state using fuzzy logic,” in 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT). IEEE, 2016, pp. 1–5