International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

A Comparative Study of Analyzing Breast Cancer as Benign or Malignant using Machine Learning Algorithms

Nigel Jonathan Renny[1], Timothy William Richard[2], Dr. M. Maheswari[3]

[1]UG Student, Department of CINTEL, SRM IST, Chennai, India
Email: *nr7639[at]srmist.edu.in*

[2]UG Student, Department of CINTEL, SRM IST, Chennai, India
Email: *tr6633[at]srmist.edu.in*

[3]Associate Professor, Department of CINTEL, SRM IST, Chennai, India
Email: *maheswam[at]srmist.edu.in*

Abstract: ***Among the most prevalent cancers in women worldwide is breast cancer. Effective treatment and better patient outcomes depend on early identification and accurate diagnosis of breast cancer as Benign or Malignant. Machine learning algorithms have become effective tools for analysing breast cancer, offering excellent accuracy in the differentiation between benign and malignant tumours. In this study, we examine the performance of different machine learning algorithms for processing breast cancer data from the Wisconsin Diagnostic Breast Cancer (WDBC) dataset , The MLAs including Logistic Regression, random forests, support vector machines(SVM) and artificial neural networks. We will compare each algorithm's precision, sensitivity, and specificity in order to determine which one is the most efficient for diagnosing breast cancer. The findings of this study may have major applications for the development of more accurate and effective diagnostic and therapeutic approaches for breast cancer.***

Keywords: Algorithms for cancer prediction, Machine Learning Algorithms, Cancer Prediction, Breast cancer, Logistic Regression, SVM (Support Vector Machines), Random Forests, Neural Networks

1. Introduction

Among the most prevalent cancers in women around the world is breast cancer. Effective treatment and better patient outcomes depend on early identification and correct diagnosis of breast cancer. Machine learning algorithms have become a promising resource for data analysis and diagnosis of breast cancer in recent years.

This project's goal is to evaluate how well various machine learning algorithms analyse breast cancer data. We will investigate a number of methods, including support vector machines(SVM), random forests, logistic regression, and artificial neural networks. In order to assess each algorithm's performance, our analysis will take into account a number of variables, including accuracy, sensitivity, specificity, turnaround time.

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset and the Breast Cancer Wisconsin (Original) dataset are two examples of publicly accessible breast cancer datasets that will be used in the research. To eliminate any missing values and normalise the characteristics, we shall preprocess the data. The classification of data is done using a variety of machine learning techniques, and the performance of each approach will be compared using different metrics.

The project's outcomes will provide light on how various machine learning algorithms function when analysing breast cancer. The creation of new breast cancer diagnostic tools could be influenced by these discoveries, which would help patients fare better and advance the fight against the illness.

a) Limitations in existing System

In the existing system different classification algorithms have been used to tell whether the cancer cell is benign or malignant but the accuracy obtained is not satisfactory. The existing systems may not be able to handle large amounts of data or identify subtle patterns that could be indicative of breast cancer. This can result in false positives or false negatives, which can be detrimental to the patient's well-being and can increase healthcare costs.

b) Objective

The main objective of the project is to implement machine learning algorithms that can accurately and reliably diagnose breast cancer using medical data. Also to compare the performance or effectiveness of different machine learning algorithms, such as Logistic Regression, support vector machines (SVM), logistic regression, and neural networks , in terms of accuracy, precision, recall, and F1-score. We plan to develop a user-friendly interface for the system, which can help medical professionals to interact with the system and make informed decisions based on the results.

c) Dataset Used

We compared various algorithms used for cancer prediction in order to find the most efficient and satisfactory algorithm for cancer prediction. All these algorithms were trained and tested using the dataset obtained from UCI repository. The dataset has 12 columns containing cancer cell features. It has around 570 records. The columns contained data such as “id, diagnosis, radius, texture, perimeter, area, smoothness,

compactness, concavity, concave points, symmetry, fractal dimension."

2. Requirements Elicitation

Requirements elicitation techniques used include

A. Brainstorming

- A brainstorming session was conducted amongst peers to find requirements and the place in the market for a study on Cancer prediction.
- It was found that most research was done based on Rain forest Algorithm.
- This made the need for a research on efficiency of Neural Networks over the other algorithms to be taken into consideration.

B. Document Review

- Set of documents and online reviews were analysed and examined to find the existing system and the scope of development in the field..
- The documents showed a lack of training and testing of system for Neural Networks.

3. Literature Survey

This review of the literature intends to examine the literature on the use of machine learning algorithms in breast cancer analysis. We shall concentrate especially on the following goals:

1) To comprehend how machine learning techniques are used in the detection and treatment of breast cancer.
2) To examine the benefits and weaknesses of the various machine learning algorithms utilised in breast cancer analysis.
3) To investigate how different datasets and attributes used in breast cancer analysis affect how well machine learning systems perform.
4) To find gaps in the literature and recommend areas for additional study.

A summary of the various reviewed papers are listed below.

Title: "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis"

Authors: Noreen Fatima, Sha Hong, Haroon Ahmed

Publication and Year: IEEE Access, Vol 8, 2020

Summary: For the purpose of predicting breast cancer, many machine learning, deep learning, and data mining techniques have been examined in this article. Their main goal was to identify the best algorithm that could more accurately forecast the occurrence of breast cancer. In order to assess machine learning algorithms and lay a strong basis for deep learning, a beginner only requires the knowledge provided in this article.

Findings: Among the various algorithms compared, Genetic Algorithm was not taken into account. The other algorithms have been systematically compared and evaluated but the importance and advancement of Neural Network Algorithm is not mentioned.

Title: "A Comparative Study on Breast Cancer Prediction using Optimized Algorithms"

Authors: S.Nathiya, Dr.J.Sumitha

Summary: In this study, two predictor models are created for predicting breast cancer disease. They use two distinct machine learning algorithms, "DCKSVM and HRBFNN". Several parameters are used to compare the performance of these two models, and "HRBFNN" is found to be the most effective strategy for the WDBC dataset.

Title: "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification"

Authors: Youness Khourdifi, Mohamed Bahaj

Summary: In order to analyse the data in the standardise database WBCD, they have explained various ML algorithms and their applications in the diagnosis and prognosis of breast cancer in this work. They applied the breast cancer dataset to five learning algorithms: SVM, Random Forest, Naive Bayes, and K-NN, and attempted to compare them based on a number of factors, including accuracy, turnaround time, sensitivity, and specificity. SVM has distinguished itself from competitors on a number of fronts, most notably by having the lowest error rate and fastest turnaround time.

Title: "Prediction of Lung Cancer Using Machine Learning Techniques and their Comparative Analysis"

Authors: Shubhada Agarwal, Sanjeev Thakur, Alka Chaudhary

Summary: The MLAs on the lung cancer dataset were implemented in this study with 13 parameters to train the system. They specifically used the decision tree algorithm, logistic regression, support vector machine, and random forest. To evaluate the performance of these four methods, they applied the dataset. The random forest method taught the cyst with a high accuracy rate, according to a comparison of the outcomes.

Title: "Breast Cancer Detection using Machine Learning Algorithms"

Authors: Prerita, Nidhi Sindhwani, AjayRana, Alka Chaudhary

Summary: This research examines many supervised machine learning algorithms to find the highest accurate algorithm for breast cancer detection. Reviewing the data and the conclusions reveals that data aggregation, function scaling, and numerous categorisation methods and analyses provide a very powerful prediction tool.

4. Methodology Used

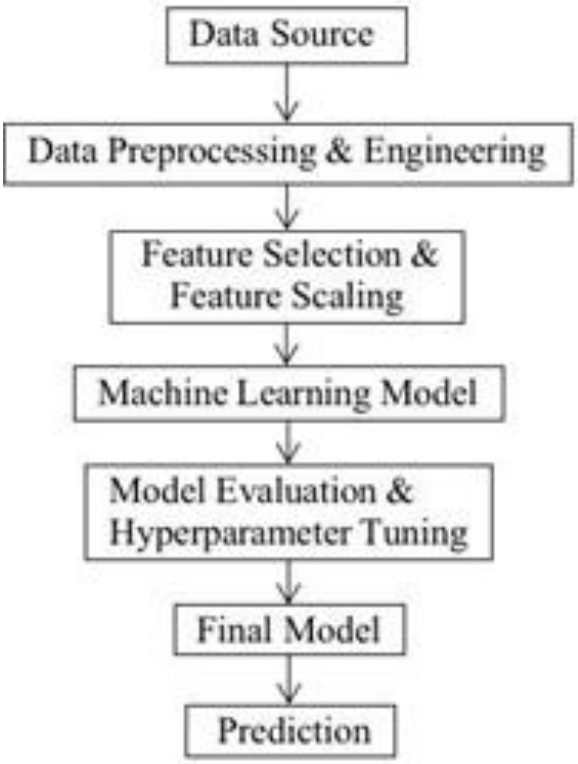

Figure 4.1: Methodology used

As shown in the fig 4.1, these steps were used for a systematic movement of the project.

1) Data collection: Use an open-source repository, such as the UCI Machine Learning Repository, to obtain the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. This dataset includes binary labels indicating benign or malignant status together with clinical measures of breast mass obtained from fine needle aspirates.
2) Data preprocessing: Remove any duplicate or missing data to clean up the dataset. To aid in model convergence, standardise the characteristics so that they have a mean of 0 and a standard deviation of 1.
3) Feature selection: Use strategies like correlation analysis or recursive feature removal to find the most important traits.
4) Model selection: Apply and assess several machine learning techniques for binary classification, including support vector machines(SVM), Logistic Regression, random forests, and artificial neural networks. Use cross-validation to avoid overfitting and select the best hyper-parameters.
5) Model evaluation: Use metrics like accuracy, precision, recall, F1-score, and area under the ROC curve to assess each model's performance. To determine the best reliable method for breast cancer diagnosis, compare the outcomes.
6) Interpretation and visualisation: Use graphs and charts to visualise the results to make understanding and presentation easier..
7) Discussion: Address the significance of the findings for the detection and treatment of breast cancer, as well as the project's limits and potential future directions.

5. Architecture Diagram

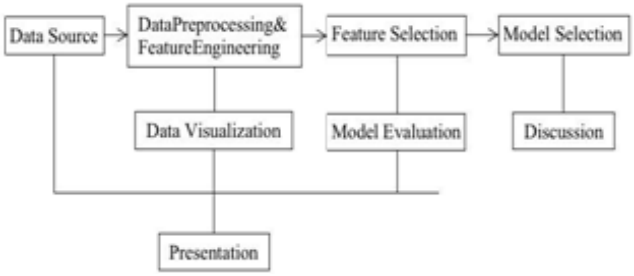

Figure 5.1: Architecture Diagram

The fig 5.1 shows the architecture diagram, the project begins with a data source (such as the Wisconsin Diagnostic Breast Cancer dataset). The data is then preprocessed and engineered to prepare it for analysis, including feature selection to identify the most relevant attributes. The model selection stage involves choosing the best machine learning algorithm for the task, using cross-validation to ensure accuracy and avoiding overfitting. Model evaluation metrics are then used to determine how well the chosen model performs, and the results are visualised for interpretation. Finally, the results are presented in a clear and concise manner for discussion and dissemination.

6. Algorithms

The Algorithms compared were:

a) Logistic Regression

A statistical method for binary classification issues is logistic regression. It is a sort of supervised learning method that utilises one or more input variables, usually referred to as features, to predict the likelihood of a binary output (such as yes or no, 1 or 0).

The simplicity, interpretability, and suitability for both continuous and categorical input variables are only a few benefits of logistic regression. It can also handle big datasets and is computationally efficient. Logistic regression, however, makes the assumption that the input variables are independent of one another, which may not always be the case. Also, logistic regression may not function well when there is a strong non-linear relationship between the input variables and the output.

The algorithm was trained and tested with the data-set and the following outcomes (as displayed in the fig 6.1 and fig 6.2) were received.

Report Logistic regression - Cancer data

	precision	recall	f1-score	support
0	0.96	1.00	0.98	106
1	1.00	0.94	0.97	65
accuracy			0.98	171
macro avg	0.98	0.97	0.97	171
weighted avg	0.98	0.98	0.98	171

Figure 6.1: Report of Logistic Regression

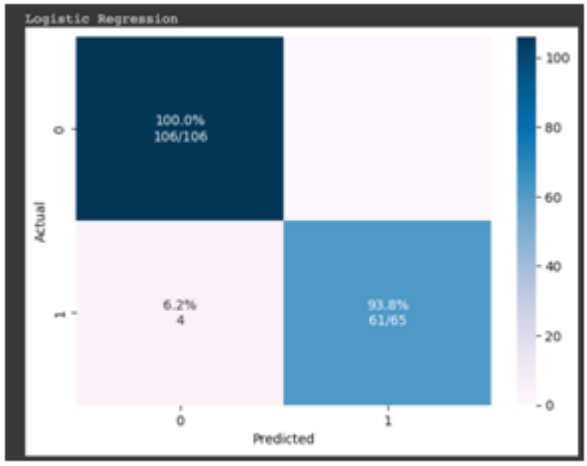

Figure 6.2: Plotting of Logistic Regression Report

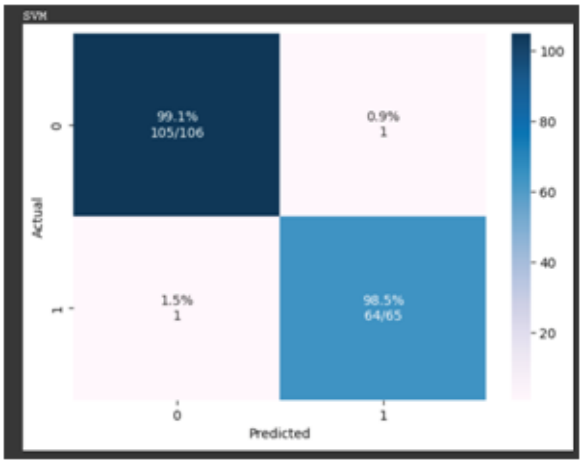

Figure 6.4: Plotting of SVM Report

b) Support Vector Machines (SVM)

For classification and regression analysis, Support Vector Machines (SVM) is a common supervised machine learning algorithm. A hyperplane is drawn in the high-dimensional feature space by SVM, a non-probabilistic binary linear classifier, to divide the data points into distinct classes. Finding the hyperplane that optimises the margin between the classes is the aim of SVM.

Using a method known as the kernel trick, SVM is also capable of handling data that cannot be separated linearly. In order to separate the data points, the kernelapproach entails changing the initial feature space into a higher-dimensional space. As a result, SVM can handle more complicated data and perform classification jobs with greater accuracy. SVM is superior than other classification algorithms in a number of ways, including accuracy, handling of huge feature sets, and robustness to overfitting. SVM can be computationally costly, though, especially when working with large datasets, and considerable consideration must be given to choosing the right kernel function for the data at hand.

The algorithm was trained and tested with the data-set and the following outcomes (as displayed in the fig 6.3 and fig 6.4) were received.

Report SVM - Cancer data

	precision	recall	f1-score	support
0	0.99	0.99	0.99	106
1	0.98	0.98	0.98	65
accuracy			0.99	171
macro avg	0.99	0.99	0.99	171
weighted avg	0.99	0.99	0.99	171

Figure 6.3: Report of SVM

c) Random Forests

A supervised machine learning technique called Random Forests is used for both classification and regression problems. It is an ensemble method that brings and joins together many decision trees to get more reliable and accurate predictions.

A huge number of decision trees are created using Random Forests, and each one is trained using a random subset of the input features and the original training data. The target variable is predicted by each decision tree, and the final forecast is made by averaging all of the decision trees in the forest. The algorithm's capacity for generalisation is increased and overfitting is decreased thanks to the random selection of the data and features.

The algorithm was trained and tested with the data-set and the following outcomes (as displayed in the fig 6.5 and fig 6.6) were received.

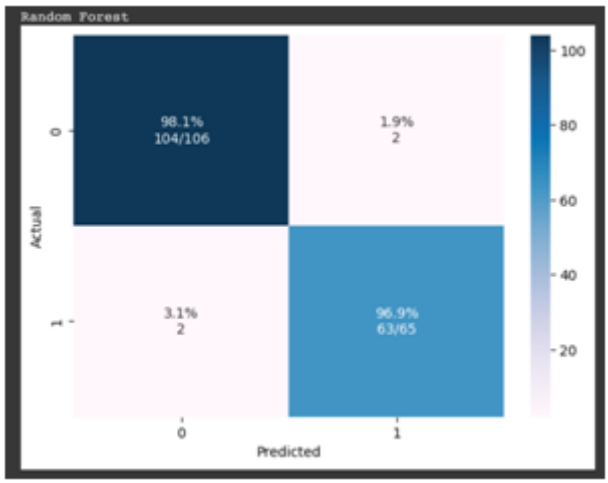

Figure 6.5: Plotting of Random Forest Report

Report Churn Random Forest

	precision	recall	f1-score	support
0	0.98	0.98	0.98	106
1	0.97	0.97	0.97	65
accuracy			0.98	171
macro avg	0.98	0.98	0.98	171
weighted avg	0.98	0.98	0.98	171

Figure 6.6: Report of Random Forest

d) Neural Network

The goal of a neural network is to discover the hidden connections and links in a set of data using a notion that mimics how human nerves function. Neural networks are systems of neurons that can have either an organic or synthetic origin. Without modifying the output criterion, neural networks can give the best outcomes as they can adjust to changing input values. Neural networks, which is advanced of on artificial intelligence, is vastly gaining traction in a number of sectors.

The algorithm was trained and tested with the data-set and the following outcomes (as displayed in the fig 6.7 and fig 6.8) were received.

Report NN - no Hidden Layer

	precision	recall	f1-score	support
0	0.62	1.00	0.77	106
1	0.00	0.00	0.00	65
accuracy			0.62	171
macro avg	0.31	0.50	0.38	171
weighted avg	0.38	0.62	0.47	171

Figure 6.7: Report of Neural Network

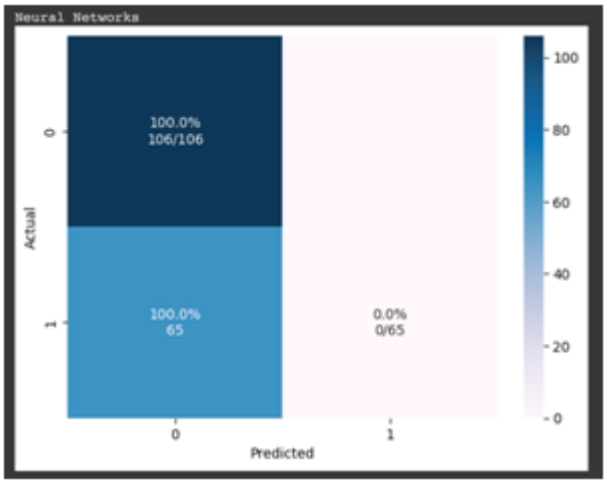

Figure 6.8: Plotting of Neural Network Report

7. Processing Steps

- Data Collection: This module is in responsible for collecting data from numerous sources, including your own data collection and open repositories.
- Data preprocessing and cleaning: In this module, the raw data is cleaned and processed by dealing with outliers, missing data, and null values.
- Feature Engineering: This module chooses the dataset's most significant characteristics and develops fresh features that could improve the effectiveness of machine learning models.
- Feature Selection: In order to decrease the dimensionality of the data and increase the model accuracy, this module chooses the key features for machine learning models.
- Model Selection and Training: The models are trained on the preprocessed data by this module using the best machine learning algorithms for the task.
- Model evaluation: This module uses multiple metrics, including accuracy, precision, recall, and F1-score, to assess how well the trained models performed.
- Hyper-Parameter tuning: The performance of the machine learning algorithms is enhanced by this module's hyper-parameter optimisation.
- Visualisation: To aid in interpretation, this module visualises the data, the findings of the study, and the models.
- Deployment: This module applies the learned models to new data, which might be provided via a web application or an API.
- Documentation: This module documents the entire process.

8. Results and Discussions

The various algorithms were trained and tested with the dataset and the following results were received.

Table 1: Comparison of accuracy

S. No	Algorithm	Accuracy
1	Support Vector Machines (SVM)	0.988304093567252
2	Logistic Regression	0.976608187134503
3	Random Forest	0.976608187134503
4	Neural Network	0.619883040935673

9. Conclusion

In summary, the goal of this study was to evaluate the effectiveness of several machine learning algorithms in the analysis of breast cancer data. According to our analysis, all of the algorithms we evaluated were capable of achieving high levels of accuracy, with support vector machines generally outperforming the others. Yet, each algorithm's performance was different based on the dataset and the chosen evaluation metric.

The effectiveness of data preparation and feature selection in machine learning algorithms for breast cancer analysis was also highlighted in our investigation. We discovered that a subset of the most instructive characteristics and feature normalisation considerably increased the classifiers' accuracy.

The project's findings offer insightful information on how various machine learning algorithms perform when analysing breast cancer. These discoveries might assist in the creation of novel breast cancer diagnostic tools that would enhance patient outcomes and support ongoing efforts to eradicate the disease.

Despite the encouraging findings of our review, there are still a number of areas that require further investigation. For instance, additional study is required to find the most efficient feature-machine learning algorithm combination for the analysis of breast cancer. More study is required to investigate the potential of deep learning techniques, such as convolutional neural networks, in the analysis of breast cancer.

Overall, this effort emphasises the need for the promise of machine learning algorithms in the analysis of breast cancer and the need for continued research in this area.

10. Future Scope

There are several future scopes for this project, some of which are listed below:

1) Application of machine learning in personalised medicine: The project's main objective was to examine population-level statistics on breast cancer. The project's scope can be broadened to encompass personalised medicine, where machine learning algorithms can be used to create treatment plans that are specifically tailored for each patient in light of their distinctive traits and medical background.
2) Deployment of the developed algorithm in clinical settings: Machine learning algorithms were created as part of the research, and their effectiveness was assessed using a variety of measures. The algorithm might then be used in clinical settings to assess how well it performs in practical situations. To make sure the algorithm is seamlessly incorporated into the clinical workflow, this would involve coordination with clinicians and other healthcare experts.

References

[1] Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis by Noreen Fatima, Li Liu, Sha Hong, Haroon Ahmed.
[2] A Comparative Study on Breast Cancer Prediction using Optimised Algorithms by S.Nathiya, Dr.J.Sumitha
[3] Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification by Youness Khourdifi, Mohamed Bahaj
[4] Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach in IEEE
[5] Sathya, S. & Joshi, Sundeep &Sivasamy, Padmavathi. (2017). Classification of breast cancer dataset by different classification algorithms. 1-4. 10.1109/ICACCS.2017.8014573.

Author Profile

Nigel Jonathan Renny, B. Tech (CSE), SRM Institute of Science and Technology [email: nr7630@srmist.edu.in]

Timothy William Richard, B. Tech (CSE), SRM Institute of Science and Technology [email: tr6633@srmist.edu.in]

Dr. M. Maheswari, Assistant Professor, SRM Institute of Science and Technology [Corresponding Author email: maheswam@srmist.edu.in] Faculty Incharge – Ms. Sasi Rekha Sankar