A COMPARISON OF ML ALGORITHMS FOR PREDICTING BREAST CANCER RISK USING HEALTH DATA

Surbhi [1], Aarti [2], Mir Junaid Rasool [3], Vinutha M [1]

[1] Department of Artificial Intelligence and Machine Learning, Acharya Institutes, Bangalore, Karnataka, India
[2] Department of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India
[3] Department of Computer Science and Engineering, Jain University, Bangalore, Karnataka, India

ABSTRACT

The detection of breast cancer is an imperative clinical task. Still, it could be a time-consuming and complicated task for medical analysis. The artificial neural network (ANN) plays an amazing part in both time optimization and accuracy, permitting the clinicals to distinguish between the patterns of malignant and benign. Here, we proposed an algorithm to find the effective methods for the breast cancer detection and classification using the ANN, random forest (RF) and support vector machine (SVM). The proposed system is composed of different stages. Firstly, preprocessing of data is done using a normalization method to remove the missing data in the dataset. At the end, classification is executed by ANN and SVM algorithms. R platform has been used for the performance evaluation of breast cancer. The results demonstrate the methodology with 99.2% accuracy using ANN architecture.

KEYWORDS

Breast Cancer Detection, SVM, Deep Learning, ANN, Random Forest, Prediction

1. INTRODUCTION

The breast cancer detection is a dynamic procedure for any medical modalities that provide the correct solution for the right time. Basically, K-means, fuzzy clustering, knowledge-based systems, ANN, SVM and expectation-maximization (EM) algorithms are the most common techniques used for the region-based segmentation to provide the necessary information from all types of medical imaging.

The number of modern technologies has been used over the decades, basically focusing on deep learning and artificial learning methodologies. The various methods such as classification, detection, facial recognition, image tracking, pattern recognition, signal processing, are examples of different areas where machine learning techniques have been applied.(Gowroju et al., 2022, Swathi et al., 2021).

Many artificial methods have been used for edge detection, discrete regional sequence, regression analysis etc. The steps to be followed for these methodologies: a) extracting the target features b)

sending the extracted features to the classifier. When it comes to detecting breast cancer, supervised learning algorithms can be effective. However, some algorithms have been found to outperform traditional supervised learning methods in many cases, especially when working with large and complex datasets.

Deep learning algorithms are part of ML that involve the use of neural networks, which are similar to the functionality of the human brain. These algorithms learn from large data, including unstructured data such as images or text. This makes them particularly useful for detecting breast cancer, which often involves analyzing large amounts of data in real-time (Swathi Gowroju et al., 2022)**,** (S Gowroju et al., 2021)

Supervised learning, involves training a model where each data point is assigned a class or category. The model is then used to predict the category of unlabeled data. While supervised learning can be effective, it may not always be able to detect complex patterns in the data that deep learning algorithms can uncover. However, based on the research conducted, deep learning algorithms have been found to outperform traditional supervised learning methods in terms of accuracy, especially when dealing with complex and large datasets.

This paper also utilizes SVM which is commonly used in both classification and regression problems. There are several reasons why SVM algorithms are preferred over other algorithms in certain scenarios: a) SVM works well in high-dimensional spaces. b) SVM is less sensitive to outliers than other algorithms, such as logistic regression, which makes it ideal for datasets with noisy data. c) SVM is capable of fitting non-linear decision boundaries using the kernel trick. This allows it to handle complex data distributions that may not be separable by a simple linear classifier. d).SVM aims to find the optimal decision boundary that maximizes the margin between different classes.

Finally, the Random Forest method has been applied, which combines different decision trees to produce accurate predictions. The trees in the forest are trained using a subset of the data, and the combined forecasts of all the trees are used to make the final prediction. Random Forest is renowned for its propensity to avoid overfitting and its capacity to handle high-dimensional data.

2. LITERATURE SURVEY

Over the past few decades, various types of CAD (Computer Aided Diagnosis) systems have been introduced, particularly in the healthcare field, utilizing diverse techniques and methods. In particular, a number of these techniques focused on utilizing data mining tools to forecast outcomes. However, in recent times, machine learning has emerged as a cutting-edge and potent discipline for uncovering patterns and making predictions based on statistical analyses, data exploration, pattern recognition, and predictive modelling.

In (Ed-Daoudy&Maalmi, 2019), A real- time health status prediction system has been created using data mining techniques, and streaming data. The system uses a machine learning model to analyze health data events and forecast health conditions, which are then saved in a distributed database. In (Michael et al., 2022), The effectiveness of the proposed framework was evaluated through the use of five different classifiers. Results from the experiment indicated that the Light-GBM classifier performed best among the other four classifiers in terms of precision, recall, F1-score and when using Bayesian optimization classifiers for 10-fold cross-validation with a tree-structured Parzen estimator. It was also found that default hyperparameter values were used in 86% of the reviewed studies. In (Amethiya et al., 2022)Two optimal techniques were identified, and the first one involved the use of the Fuzzy ELM-RBF ML algorithm, which yielded the

highest accuracy of 98.05% for detecting breast cancer. (Rasool et al., 2020) built a comprehensive prediction framework around Spark, utilizing machine learning, SQL Spark and Spark streaming. The models were developed to determine the best classifier in terms of performance, and found that random forest was the most effective for the use case dataset. (Vaka et al., 2020) A new approach, DNNS, has been developed for detecting breast cancer, which sets itself apart from other methods by utilizingValue of Support on a deep neural network. To further improve performance, accuracy, and quality of image, a normalizationhas been integrated. Experimental results indicate that the DNNS approach surpasses current techniques, providing benefits in terms of both performance and image quality, which are critical factors in modern medical systems. (Nemade et al., 2023) In this work despite the development of numerous methods for breast cancer classification, achieving high accuracy remains a challenge. To address this issue, work proposed a model for breast cancer classification and compared six machine learning classification techniques, along with ensemble techniques, on thedataset. Findings showed that the DT classifier with a gini index criterion achieved the highest accuracy of 97% among the standard ML algorithms, and the LR classifier achieved the highest AUC of 0.996. (Rasool, 2022) aims to develop a machine learning model using Spark's MLlib to predict health status. The system utilizes stroke data with specific attributes to train machine learning models for disease prediction. ML algorithms applied to the dataset, and their performance is compared using a confusion matrix to calculate accuracy.(Rasool, 2020) For the offline model, four classifiers - were tested to determine the best one for the final monitoring system. To evaluate their performance, the heart disease dataset was split into 70% training and 30% testing data, and accuracy and other parameters were calculated. The algorithms used different arguments, such as maxIter and regParam for LR, maxIter for GBT, and maxDepth, maxBins, and numTrees for DTC and RF.

3. EVALUATION CRITERIA

The effectiveness of the classification model is calculated using a variety of evaluation indicators. The metrics, which include specificity, sensitivity, and accuracy, are computed for the predicted class and the actual class. The degree to which a model can accurately detect negative cases is measured by its specificity. In more detail, it is the ratio of actual negative cases that the model correctly predicts as negative.

$$Specificity = \frac{TN}{(FP+TN)} \qquad (1)$$

Sensitivity is a metric that evaluates the accuracy of optimistic predictions. The proportion of True Positive to the total number of actual positive instances is used to determine it. It can also be stated as the proportion of True Positive to the total of True Positive and False Negative.

$$Sensitivity = \frac{TP}{(FN+TP)} \qquad (2)$$

The overall correctness of the model's predictions is measured by the accuracy metric. It is determined by the ratio of the sum of True Positive and True Negative with the total number of instances.

$$Accuracy = \frac{(TN+TP)}{(TN+FP+FN+TP)} \qquad (3)$$

We can evaluate the performance of model for two classes, the actual class, and the predicted class, using the confusion matrix and evaluating specificity, sensitivity, and accuracy.

4. METHODOLOGY

Different steps are carried out in a particular order to fulfil the proposed objective as explained in figure 2. Three different algorithms are used to build the model and specificity, accuracy, and sensitivity is used to calculate model performance. The model comparison is done on the basis of parameters like KS(Youden), Gini, Lift, and Misclassification. Out of the three algorithms i.e. SVM, RF, and ANN the one with the highest accuracy is selected.

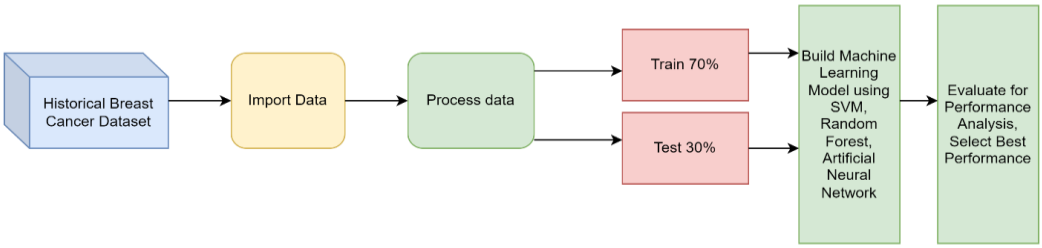

Figure 2: Architecture of predicted system

a) Programming Language used :R programming is used for implementing the proposed work in Rstudio. R is one of the best-known languages for statistics and machine learning. It has many inbuilt packages and libraries for the implementation of machine learning algorithms.

b) Data Collection: UCI Repository's (http://archive.ics.uci.edu/ml) Wisconsin Breast Cancer Diagnostic dataset is selected for the proposed work. It includes 569 observations of cancer biopsies, with 32 variables, including an id, diagnosis, and 30 numeric measurements representing cell nuclei characteristics from digitized images. These 30 variables of standard error, mean, and worst value for 10 different characteristics. The diagnosis is coded as "M" for malignant (Harmful) and "B" for benign(Not Harmful).

c) Data Exploration and Preparation : The variable diagnosis is selected as the class attribute with two values i.e. "M" and "B" which we have to predict. The class variable is converted to factor data for the implementation. As id is the unique identifier for each patient and does not provide any useful information, it is dropped from the model. For rescaling the data to the standard range of values, all the numeric variables are normalized. The data extraction is applied to partition data into training(70%) and testing(30%) datasets.

d) Model Training : The training dataset is used for building the model. Three different models are built on the same data. The algorithms are implemented using inbuilt packages and libraries present in R. Some of the R libraries and packages used in the proposed system are as follows:

- **Neuralnet:** The neuralnet package by Fritsch and Guenther is selected for the implementation of ANN. For training the model neuralnet() function is used which is present in the same package.
- **Kernlab:** For the implementation of SVM, kernlab package is used as it is better than others present. The function used for training SVM classifier is ksvm().
- **Random Forest:** The classifier is built using randomForest() present in RandomForest package of R. It is classification and regression based on a forest of trees using random inputs created by Breiman and Cutler.

For each model built specific parametric values are selected and three models are created. For SVM, cubic kernel is used. The splitting criteria for Random Forest is Chi-square. Hyperbolic Tangent is selected as Activation function.

e) Model Testing The models built on the training dataset is then applied on the testing dataset to predict the values of the class attribute. The predicted values and actual values are compared of the same variable of testing dataset to form the confusion matrix.

f) Comparing Model Performance The models built are compared on three criterias Lift, Gini, and Misclassification. These criteria are commonly used in machine learning to measure the accuracy and usefulness of a model. The fit statistic, KS(Youden), which measures how well the model distinguishes between positive and negative cases and considers both sensitivity and specificity. Their values are compared and best model is selected.

5. RESULTS

The breast cancer data is first explored, and pre-processing is performed in which data is cleaned and normalized which is then divided into testing (30%) and training(70%) datasets. Three algorithms are used to build three different prediction models on training dataset and the performance is checked on testing one. The confusion matrix values for each of the algorithm is shown in Table 1.

Table 1 Confusion Matrix of SVM, Random Forest and ANN

	SVM		RF		ANN	
	M	B	M	B	M	B
M	205	7	205	7	208	4
B	0	357	3	354	0	357

The results of Specificity, Accuracy and Sensitivity for each algorithm on different parameters is shown in table 2. The highest accuracy attained is 99.2% for Artificial Neural Network making it the best performing algorithm in detecting breast cancer.

Table 2 Specificity, Sensitivity and Accuracy of three algorithms

Algorithm	**Specificity**	**Sensitivity**	**Accuracy**
Support Vector Machine Kernal: Cubic Penalty Value: 1 Tolerance Value: 0.000001 Maximum Iterations: 25	3.30	100	98.7
Random Forest Number of trees: 100 Bootstrap: 0.6 Splitting Criterion: Chi- square Number of predictors to split node: 6	3.30	99.1	98.2
Artificial Neural Network Maximum Iterations: 250 Optimization methos: LBFGS L1: 0 L2: 0.1 Hidden layers: 2 Neurons: 10 per hidden layer Activation Function: Hyperbolic Tangent for each hidden layer	1.88	100	99.2

Table 3 shows the performance of each model using different metrices. For the assessment, number of bins is 20 and tolerance is 1e-10. The value of lift for all the three algorithms is same and the misclassifications done by ANN is lesser than that of SVM and Random Forest. The value of KS(Youden) and Gini for ANN is higher than the other two algorithms. Also, all these assessment metrices also depicts that the best performer is Artificial Neural Network.

Table 3: Performance Assessment of algorithms

Algorithm	**KS(Youden)**	**Lift**	**Gini**	**Misclassification**
Support Vector Machine	0.977	2.7358	0.997	0.012
Random Forest	0.959	2.7358	0.994	0.018
Artificial Neural Network	0.989	2.7358	1	0.007

Figure 2 shows the comparison graph of SVM, RF and ANN using different statistical metrices. This shows that the ANN is best for detecting breast cancer in patients than SVM and RF.

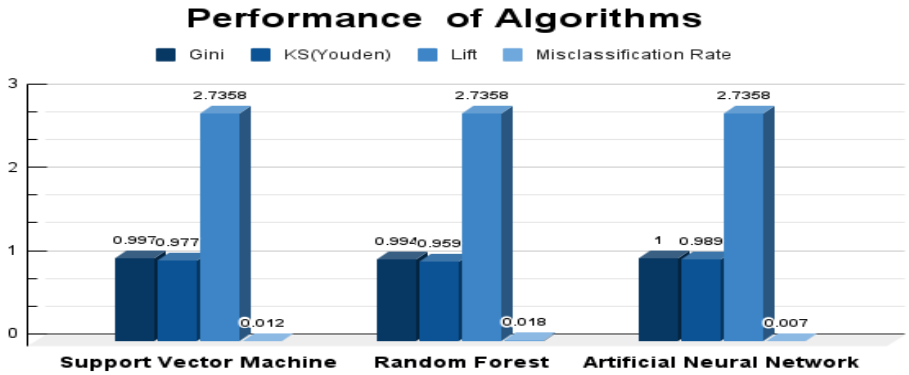

Figure 2 Performance Comparison Graph of algorithms

6. CONCLUSION

In this research we have built three different models to check their accuracy in detecting breast cancer using three main algorithms, SVM, Random Forest and ANN. On testing the model performance on different metrices we have concluded that Artificial Neural Network proves to be best among the algorithms under consideration on the selected dataset. The majority of women duffers from breast cancer and it can be cured if detected on time. In the conventional system the disease detection takes a lot more time but with the machine learning model it can be done easily in less time after getting the different attribute values to predict whether patient has breast cancer. The model comparison done in this paper concluded that ANN with 99.2% accuracy can be used in timely detecting of the disease.

REFERENCES

[1] Amethiya, Y., Pipariya, P., Patel, S., & Shah, M. (2022). Comparative analysis of breast cancer detection using machine learning and biosensors. *Intelligent Medicine*, *2*(2), 69–81. https://doi.org/10.1016/j.imed.2021.08.004

[2] Ed-Daoudy, A., &Maalmi, K. (2019). Application of Machine Learning Model on Streaming Health Data Event in Real-Time to Predict Health Status Using Spark. *International Symposium on Advanced Electrical and Communication Technologies, ISAECT 2018 - Proceedings*, 1–4. https://doi.org/10.1109/ISAECT.2018.8618860

[3] Kabiraj, S., Raihan, M., Alvi, N., Afrin, M., Akter, L., Sohagi, S. A., & Podder, E. (2020). Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm. *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*, 11–14. https://doi.org/10.1109/ICCCNT49239.2020.9225451

[4] Michael, E., Ma, H., Li, H., & Qi, S. (2022). An Optimized Framework for Breast Cancer Classification Using Machine Learning. *BioMed Research International*, *2022*. https://doi.org/10.1155/2022/8482022

[5] Nemade, V., Fegade, V., Nemade, V., &Fegade, V. (2023). ScienceDirect ScienceDirect Machine Learning Techniques for Breast Cancer Prediction Machine Learning Techniques for Breast Cancer Prediction. *Procedia Computer Science*, *218*(2022), 1314–1320. https://doi.org/10.1016/j.procs.2023.01.110

[6] Rasool, M. J. (2020). Spark based Health Status Prediction on Real Time Streaming Data using Machine Learning. *International Journal for Research in Applied Science and Engineering Technology*, *8*(11), 471–478. https://doi.org/10.22214/ijraset.2020.32197

[7] Rasool, M. J. (2022). *Machine Learning on Healthcare Big Data Using Apache Spark Machine Learning on Healthcare Big Data Using Apache Spark*. *March*.

[8] Rasool, M. J., Brar, A. S., & Kang, H. S. (2020). *Risk Prediction of Breast Cancer From Real Time Streaming Health Data Using Machine Learning*. *11*, 409–418. https://doi.org/10.5281/zenodo.4284315

[9] Vaka, A. R., Soni, B., & K., S. R. (2020). Breast cancer detection by leveraging Machine Learning. *ICT Express*, *6*(4), 320–324. https://doi.org/10.1016/j.icte.2020.04.009

[10] Gowroju, S, Aarti, & Kumar, S. (2021). Robust Pupil Segmentation using UNET and Morphological Image Processing. *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, 105–109. https://doi.org/10.1109/MIUCC52538.2021.9447658

[11] Gowroju, Swathi, Kumar, S., Aarti, & Ghimire, A. (2022). Deep Neural Network for Accurate Age Group Prediction through Pupil Using the Optimized UNet Model. *Mathematical Problems in Engineering*, *2022*, 7813701. https://doi.org/10.1155/2022/7813701

[12] Swathi, A., Aarti, & Kumar, S. (2021). A smart application to detect pupil for small dataset with low illumination. *Innovations in Systems and Software Engineering*, *17*(1), 29–43. https://doi.org/10.1007/s11334-020-00382-3

[13] Aarti, Sikka, G. and Dhir, (2020) Novel Grey Relational Feature Extraction Algorithm for Software Fault-Proneness Using BBO (B-GRA). Arabian Journal for Science and Engineering, 45(4), 2645–2662.

[14] Aarti, Sikka, G. and Dhir, R. (2019) Empirical validation of object-oriented metrics on cross-projects with different severity levels. International Journal of Computational Systems Engineering (IJCSYSE), Inderscience, 5(5/6), 304-332

[15] Aarti, Sikka, G. and Dhir, R. (2018) Grey relational classification algorithm for software fault-proneness with SOM clustering. International Journal of Data Mining, Modelling and Management (IJDMMM), Inderscience, 12(1), 28-64