JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Comparison of Various Machine Learning Algorithms for the Detection of Breast Cancer

Raja Tawseef Ahmad Mir

[1]M. Tech Scholar, Universal Group of Institutions,Lalru,Punjab India

Preeti Sondhi

Assistant Professor, Universal Group of Institutions,Lalru,Punjab India

***Abstract*:** Breast cancer is the most common disease in women and a significant factor in the growing death rate for women. Given that manual breast cancer diagnosis requires a lot of time and that there aren't many methods accessible, there is a great demand for automation diagnostic tools for early detection of breast cancer. The development of such systems heavily relies on deep learning and machine learning techniques. For the purpose of differentiating between benign and malignant tumors, we have used machine learning classification methods. These methods enable the computer to learn from the past data and anticipate the type of new input. Breast cancer is the most prevalent cancer in women (43.3 instances per 100,000 women) and the one with the highest mortality rate (14.3 incidents per 100.000 women). Early diagnosis is crucial for survival. Machine learning approaches may be used to successfully identify, predict, and assess the problem.

The first part of this work is to present the datase, what it contains, when and how it was created, if it is noisy, if it has missing values. This section is important to understand what are the issues that will need to be processed while preparing the data to create the classifier. The next step is to propose methods and algorithms to optimize the training set. How to deal with missing values? How to avoid overfitting the classifier? All these questions are discussed and different solutions are proposed.

The Gaussian Naive Bayes (GNB), k Nearest Neighbours (K-NN), Support Vector Machine (SVM), Random Forest (RF), AdaBoost, Gradient Boosting (GB), XG Boost, and Multi-Layer Perceptron were the eight machine learning techniques that we examined in this study (MLP). The experiment uses datasets from Breast Cancer Wisconsin together with a confusion matrix and 5-fold cross-validation. The results of the tests showed that XGBoost had the highest performance. XGBoost was used to achieve accuracy (97,19%), recall (96,75%), precision (97,28%), F1-score (96,99%), and AUC (99,61 percent). Our results showed that the best method for predicting breast cancer in the Breast Cancer Wisconsin dataset is XGBoost.

Keywords: XG Boost, Breast Cancer, Machine learning , Random Forest

1 INTRODUCTION

One of the top causes of death for women is breast cancer (after lung cancer). In the US, it is expected that women will get breast cancer 246,660 times in 2016, and that 40,450 women will pass away from the disease. 25% of all malignancies in women and around 12% of all new cases of cancer are breast cancer. [1] There may be uses for machine learning and deep learning in the treatment of cancer. The terms "machine learning" and "deep learning," which are now used interchangeably with "artificial intelligence" and "data science," have fundamentally altered how choices are made and results are predicted using artificial intelligence.. In actuality, big data has increased both the amount of data and the ability to extract value from it. Machine learning & deep learning technologies, for instance, are being utilized to solve medical science problems more and more frequently because to their excellent performance in outcome prediction, saving medical costs, and increasing patients' health while making judgments in real time to save lives. Breast cancer is the most common cancer in women, with 43.3 incidences per 100,000 women. [2] Breast cancer has a fairly low fatality rate when compared to other malignancies. However, since there are so many cases, it has the highest mortality rate of any cancer in women (12.9 per 100 000).

As the dataset of breast cancer patients expands, machine learning techniques are more likely to be employed to give a quick, automated, and improved understanding of cancer healthcare (Maity, G., and Das, S. 2017). The vast databases that are easily accessible for detection provide us the opportunity to generate an accurate prognosis. The problem is figuring out which strategy will result in the best result.[4]

Earlier studies looked into several machine learning strategies for predicting breast cancer. Using the Wisconsin Breast Cancer dataset, Asri, H. et al. examined the effectiveness of Support Vector Machines (SVM), Decision Tree (C4.5), Nave Bayes, and k-Nearest Network (kNN) (Asri, H. et al. 2016).[5] Compared to other methods, SVM performed better and had the highest accuracy (97,13%). Bayrak, E. et al. (Bayrak, E. et al. 2019) tested SVM with Artificial Neural Network to predict breast cancer in its early stages (ANN). The findings showed that SVM performed the best, with a score of 96,9957 percent. Gbenga, D. et al. [5] investigated eight machine learning methods to identify breast cancer using the WEKA data mining and machine learning

simulation environment. The methods that were compared in this study were SVM, Radial Based Function, Simple Linear Logistic Regression Model, Naive Bayes, kNN, AdaBoost, Fuzzy Unordered Role Induction approach, and Decision Tree (J48). Their experiment's findings demonstrated that SVM performed at its peak (97.07 percent)

Various results show that a comparison of these approaches is doable. A common dataset may be used to objectively determine the best strategy. There are more, never-before-comparable methods for predicting breast cancer. We evaluated the performance of the following machine learning methods as a result: AdaBoost, Gradient Boosting (GB), XGBoost, K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), and Gaussian Naive Bayes (GNB).

Dataset

The dataset was downloaded from Kaggle and comprises of two classes, Healthy and Sick. Each class comprises of 50 images and is of Magnetic resonance imaging nature.

2 MOTIVATION AND OBJECTIVES

Early diagnosis is crucial for survival. In low- and middle-income countries, cancer mortality account for around 70% of all deaths. Patients find it difficult to get an appointment with a doctor due to a lack of funds and underdeveloped healthcare facilities. By developing early diagnostic programs based on early signs and symptoms, the patient survival rate can be increased. World Health Organization; 2019 [3]. In this work , the following objectives are intended to be achieved

1. Researching breast cancer detection algorithms such KNN, XG Boost, Adaboost, SVM, and Random Forest

2. To use the Anaconda and Jupiter tools to apply the models to the dataset.

3. To compare the outcomes of various model-specific factors, such as Epochs and Batch Size.

4. To assess the outcomes and determine the optimum model.

3 LITERATURE REVIEW

Using decision tree classifier (CART) technology, Angeline Christopher. Y and Dr Sivaprakasam98 achieve accuracy of 69.23% in datasets related to cancer. Pradesh compares the efficacy of the machine learning algorithms SVM, NN, and BF Tree. SMO's performance is superior to that of other classifiers, according to the data. When utilizing Wisconsin cancer (original) datasets, Joe achieves an accuracy of 95.06%. In this study, a hybrid approach was suggested to improve the Wisconsin carcinoma (original) datasets' classification accuracy (95.96), using 10 fold cross validation..

In order to predict carcinoma survival and growth, Liu Ya-Qin, W. Cheng, and Z. Lu18 experimented on carcinoma data using the C5 algorithm with bagging. They did this by providing substantial dataset to train from the first set using variations with cycles to supply multiple sets of data that are the same size as the first data. The record of 202,932 cancer patients are taken by Delen89 et al. Lu19 and before the into two categories: "survived" (93,273) and "not survived" (109,659). The accuracy of the results for individually to ensure was around 93%. referring to all of the previously listed linked work. My research evaluates how the machine learning algorithms SVM, DT, k-NN and NN, Logistic regression, XG-Boost, and RF behave in the diagnosis and analysis of Wisconsin cancer (original) datasets to make judgments. The objective is to analyze data with the utmost simplicity and the lowest possible mistake rate. I evaluate the usefulness and efficiency of these methods using a variety of metrics, such as accuracy, instances that were properly and wrongly categorized, and the amount of time it took to develop the model, among others.

4 METHODOLOGY

Deep learning is a branch of machine learning that makes use of multi-layered neural networks to give computers the ability to imitate human learning. Deep learning enables voice control in consumer electronics like smart phones, tablets, and other smart devices. It teaches autonomous cars to distinguish between a pedestrian and a tree or to recognize a road sign.

A deep learning model may be trained to carry out classification tasks directly using input in the form of examples, such as text, images, or audio. In terms of accuracy, deep learning algorithms routinely outperform human performance. They are trained using a sizable labeled dataset.

Although deep learning was first postulated in the 1980s, it has only recently proven beneficial for the following reasons:

why it was impossible to collect the quantity of labeled data required for deep learning at the time. For instance, the development of driverless automobiles requires the successful face detection of millions of images and countless hours of video.

Strong computers are required for deep learning. Rather of taking weeks or months to train, the network can now do it in only a few hours thanks to modern, strong GPUs and parallel computations supported by clusters or cloud computing.

Since the bulk of deep learning approaches use neural network topologies, fully connected layers are occasionally referred to as "deep neural networks".

The term "deep" implies that there are several hidden layers in the neural network. In contrast to the standard 2-3 hidden layers seen in neural networks like perceptions, deep networks may have up to 150 hidden layers.

These models were trained on big labeled datasets using architectures that allow learning parts directly out from data alone without requirement for manual extracting the features.

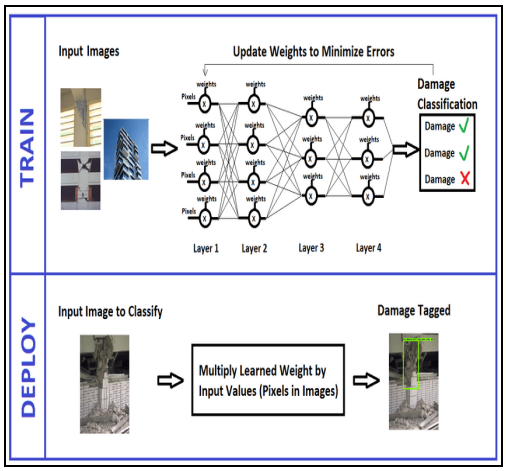


Figure 4.1: Deep Learning approach

4.1 Machine Learning VS Deep Learning comparison

Deep learning is a part of machine learning. Using pertinent attributes that were manually gathered as part of the standard machine learning process, a model is developed that categorizes the examples' elements.

In a deep learning process, pertinent properties are automatically obtained from instances. Deep learning also uses "end-to-end learning," when a model is given a task, such as classification, and then learns how to carry it out on its own.

While machine learning programs convergence at shallow learning, deep learning techniques grow with data. Shallow learning suggests that after adding a specific quantity of training data to the network, algorithms reach a performance plateau at a specific level.

As data size grows, networks for deep learning should continue to advance.

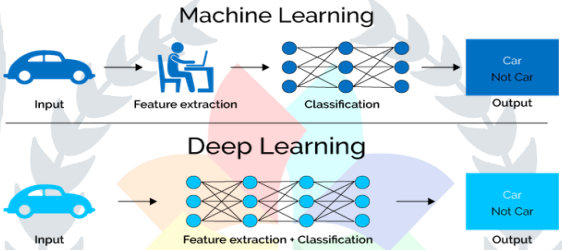


Figure 4.1: Comparison: ML and DL

4.2 ML algorithms and methods

4.2.1 Decision Tree Algorithm:

A straightforward visual aid for categorizing samples is a decision tree. It is a type of supervised machine learning in which the data is continually divided based on a certain feature.

Decision Tree includes:

Test the value of a certain characteristic with nodes.

• Edges/Branches: Relate to the subsequent node or leaf based on the results of a test.

• Leaf nodes: Station nodes that convey labels or class distribution and anticipate the result

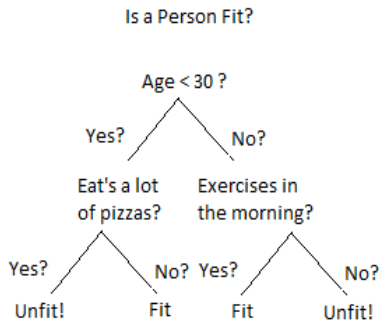


Figure 4.2: Decision Tree example

The Random Forest construction pseudo-code is as follows:

• As a first measure, we divide the data according to the characteristic that produces the greatest classification algorithm (IG) (reduction in uncertainty towards the final decision).

• We may then resume this breaking process at each child node in an iterative manner until all of the examples at each leaf node fall into the same class.

• In order to avoid over-fitting, in practice we can impose a restriction on the depth of the tree. As the finished leaves may still contain certain impurities, we make a slight tradeoff in terms of cleanliness.

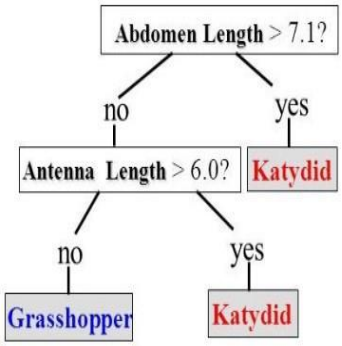


Figure 4.3: Other Decision Tree example

The decision tree classifier has the following benefits:

• The structure of the little tree is simple to explain in terms of human knowledge.

• For small data sets, it is accurate compared to other classification methods.

• Easy to assemble

• Very quick to analyze brand-new (unknown) data points.

• Ignores inconsequential details.

4.2.2 Random Forest Algorithm:

A supervised classification technique is the Random Forest algorithm. As previously said, it is built on decision trees. It will be made random and will consist of a forest of decision trees. There is a correlation between the outcomes and the number of decision trees in the forest; the more decision trees, the more realistic the results. But it's important to remember that building the forest is different from building the decision trees using information gain or other metrics.

There are four key benefits of random forest algorithms. Both classification and regression issues may be solved with it. However, if there are enough trees in the forest, the classifier won't over-fit the model, which might make the results worse. The last benefit of using Random Forest is that it can be used to create classifiers for categorical data rather than numeric data.

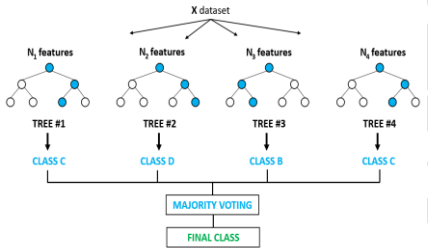


Figure 4.4: Random Forest example

The Random Forest construction pseudo-code is as follows:

(1) Choosing "K" features at random from all "m" features, where "k" > "m"

(2) Use the optimal breaking point to determine the node "d" from the "K" characteristics.

(3) Use the best break to split the node into daughter nodes.

(4) Continue with steps 1-3 until "l" nodes are reached.

(5) Build a forest by repeating steps 1 through 4 "n" times to create "n" trees.

When compared to other classifier, the Random Forest classifier has the clear advantage:

• It avoids over-fitting for classification problems

• It can be used to extract the most essential features from the dataset, providing information for other techniques for machine learning.

The same approach may be used to issues involving regression analysis and classification.

5 SYSTEM IMPLEMENTATION

This work makes use of the Wisconsin Breast Cancer (original) datasets from the UCI Machine Learning Repository. 569 cases (350 benign and 219 malignant), 2 classes (65.5% malignant and 34.5% benign), and 11 integer-valued characteristics are included in the dataset. Following figure gives the flow diagram of the proposed system

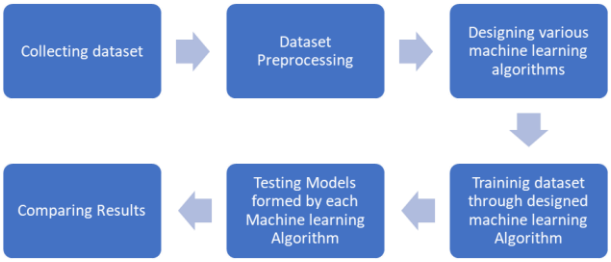


Figure 5.1 Flow Diagram of the System

5.1 Collecting data set

The dataset was downloaded from Kaggle and comprises of two classes, Healthy and Sick. Each class comprises of 50 images and is of Magnetic resonance imaging nature. The data set that we chose to work on was Diagnostic Wisconsin breast cancer Database which is a widely used and worldwide accepted dataset and easy to use and find.

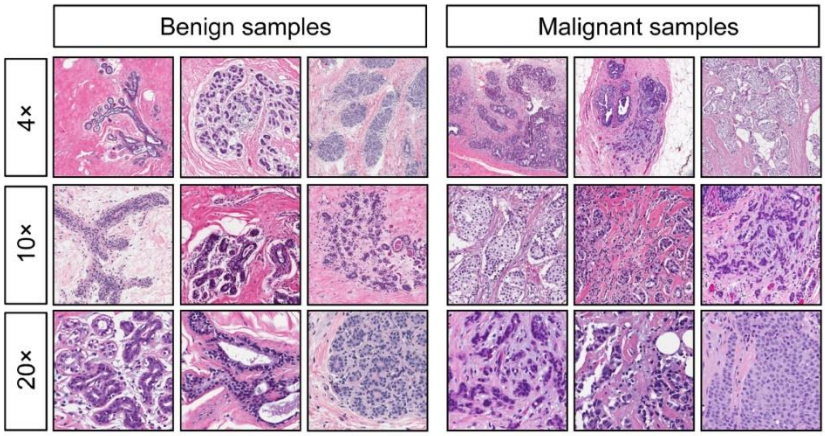


Figure 5.2 Sample Dataset

5.2 Dataset preprocessing

The "Diagnostic Wisconsin breast cancer Database" has a variety of characteristics. These qualities include:

1) ID

2) The "diagnosis" (B = benign; M = malignant)

3) For each cell nucleus, ten characteristics will be calculated, including:

• Radar (mean of distances from center to points on the perimeter)

The texture (standard deviation of gray-scale values)

Area

• Perimeter

• Easiness (local variation in radius lengths)

• Compactness (area - 1.0 / perimeter 2)

• Curvature (severity of concave portions of the contour)

• Concave edges (number of concave portions of the contour)

• Fractal dimension; symmetry ("coastline approximation" - 1)

4) The three sections of attributes 3-32 are as follows:

The worst, Stranded-Error (13-23), Mean (3-13), and (23-32)

Radius, texture, area, width, creaminess, compactness, convexity, concave points, symmetry, and fractal dimension are the 10 parameters that each model has.

• Mean (The mean of the total cells) (The mean of the all cells)

• Common Error (The standard Error of all cell)

• Stupidest (The mean of worst cells)

5.3 Designing various machine learning modules

Various machine design modules were designed for the research and comparison purposes which include SVM, AdaBoost, Random Forest, KNN, XG Boost, Grabbing, and ANN. The design parameters, including training Accuracy, Testing Accuracy, Testing Data, F1 score, Recall, and Precision, were saved and stored.

5.4 Training

The data set was trained over these modules and thoroughly observed for the errors and complications where too many false-positive results were taken care of from the algorithm in order to accurately diagnose breast cancer. For instance, false-negatives or (Anyone don't have cancer, but we told them to get the therapy). That is the primary reason the model with the best overall accuracy is picked.

Figure 5.3 Training results

diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
M	17.99	10.38	122.80	1001.0	0.11840	0.27760
M	20.57	17.77	132.90	1326.0	0.08474	0.07864
M	19.69	21.25	130.00	1203.0	0.10960	0.15990
M	11.42	20.38	77.58	386.1	0.14250	0.28390
M	20.29	14.34	135.10	1297.0	0.10030	0.13280

Figure 5.4 Parameters of the detected tumours over the dataset

compactness_mean	concavity_mean	concave points_mean	...	radius_worst	texture_worst	perimeter_worst	area_worst
0.27760	0.3001	0.14710	...	25.38	17.33	184.60	2019.0
0.07864	0.0869	0.07017	...	24.99	23.41	158.80	1956.0
0.15990	0.1974	0.12790	...	23.57	25.53	152.50	1709.0
0.28390	0.2414	0.10520	...	14.91	26.50	98.87	567.7
0.13280	0.1980	0.10430	...	22.54	16.67	152.20	1575.0

Figure 5.5 Various measurements of the tumour

...rst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst
	0.1622	0.6656	0.7119	0.2654	0.4601	0.11890
	0.1238	0.1866	0.2416	0.1860	0.2750	0.08902
	0.1444	0.4245	0.4504	0.2430	0.3613	0.08758
	0.2098	0.8663	0.6869	0.2575	0.6638	0.17300
	0.1374	0.2050	0.4000	0.1625	0.2364	0.07678

Figure 5.6 various measurements of the cancers detected on the dataset

5.5 Testing

Testing was done over all the algorithms and the accuracy was tracked and noted for further comparisons

5.6 Co-Relation Map between attributes

Figure 5.7 Correleation over the model

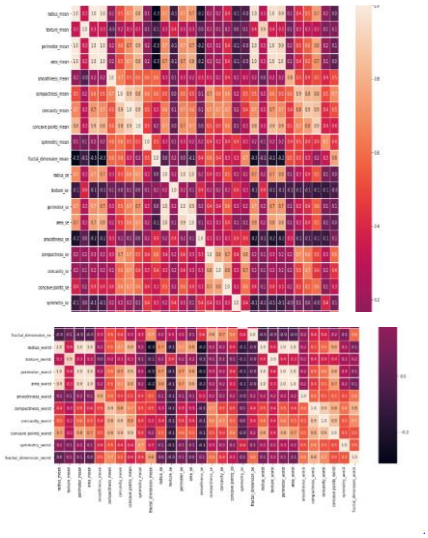
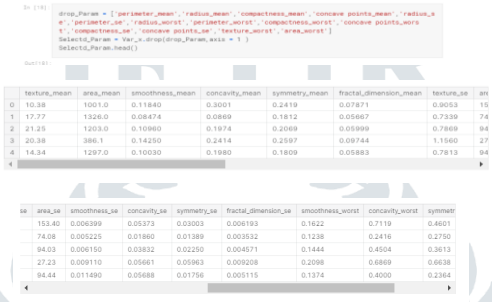


Figure 5.8 Feature relation map

5.7 Selected parameters for training the model



5.8 Comparison of the results

It was discovered that ANN surpasses all other models in every parameter of the comparison study after viewing and contrasting all the models under all the various parameters. The metrics that were evaluated as being the greatest and the best were Training Accuracy=1.0, Testing Accuracy=1.0, Validation Accuracy=1.0, F1 score=0.999, Recall=0.993, and Precision=1.0.

6 SIMULATION AND RESULTS

The training dataset count was gradually increased from 0.1 to 1.0 from all the comparative analysis of 7 models, including SVM, AdaBoost, Random Forest, KNN, XG Boost, Grabbing, and ANN. The design parameters, including training Accuracy, Testing Accuracy, Testing Data, F1 score, Recall, and Precision, were saved and stored.

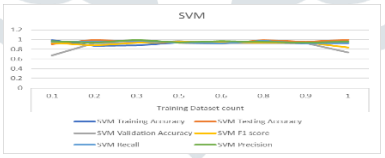


Figure 6.1 SVM Model Analysis

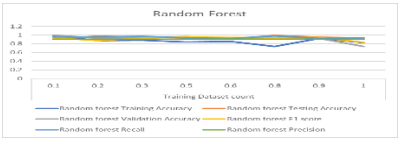


Figure 6.2 Random Forest Model Analysis

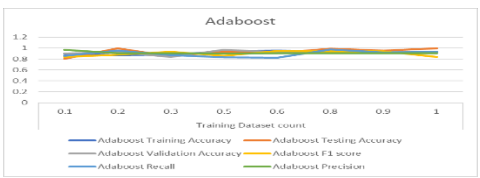


Figure 6.3 Adaboost Model Analysis

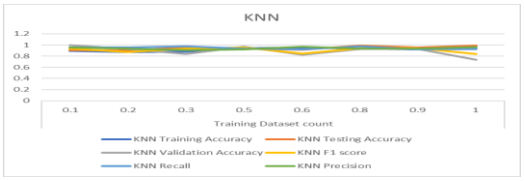


Figure 6.4 KNN Model Analysis

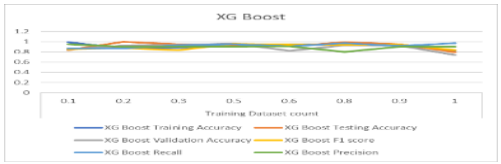


Figure 6.5 XG Boost Model Analysis

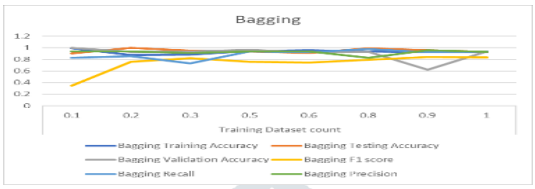


Figure 6.6 Bagging Model Analysis

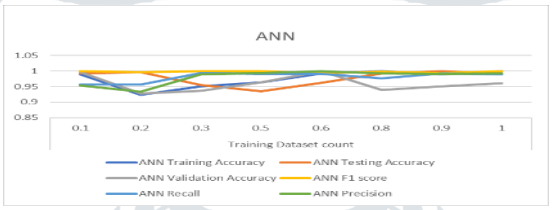


Figure 6.7 ANN Model Analysis

6.1 Comparative Analysis

The best training accuracy for the SVM model was 0.9916, the highest testing accuracy was 0.97, the highest validation accuracy was 0.92, the F1 score was 0.94, the Recall was 0.97, and the least Precision was 0.925, according to the analysis.

The following values were found for random forest: F1 score = 0.93, Recall = 0.97, Precision = 0.903, Training accuracy = 0.99, Testing accuracy = 0.97, Validation accuracy = 0.96.

The best and highest values for AdaBoost's criteria were Training accuracy = 0.95, Testing accuracy = 0.99, Verification accuracy = 0.96, f1 score = 0.945, Recall = 0.93, and Precision = 0.9026.

It was found that for KNN, Training Accuracy = 0.95 and Testing Accuracy = 0.95. Validation Precision = 0.9 F1 rating: 0.97 Sophistication: 0.905, Recall: 0.97

Training Accuracy for XG boost was 0.99, Testing Accuracy was 0.93, Validation Accuracy was 0.96, F1 score was 0.95, Recall was 0.97, and Precision was 0.94.

Training Accuracy was 0.99, Testing Accuracy was 0.95, Validation Accuracy was 0.99, F1 score was 0.85, Recall was 0.97, and Precision was 0.99 for baggage.

It was discovered that ANN surpasses all other models in every parameter of the comparison study after viewing and contrasting all the models under all the various parameters. The metrics that were evaluated as being the greatest and the best were Training Accuracy=1.0, Testing Accuracy=1.0, Validation Accuracy=1.0, F1 score=0.999, Recall=0.993, and Precision=1.0.

Table 6.1: Comparison table

	SVM	Random Forest	Ada Boost	KNN	XG Boost	Baging	ANN
Training Accuracy	0.99	0.97	0.99	0.95	0.99	0.99	1
Testing Accuracy	0.97	0.96	0.96	0.95	0.93	0.95	1
Validation Accuracy	0.92	0.96	0.94	0.9	96	0.99	0.98
F1 score	0.94	0.93	0.93	0.97	0.95	0.85	0.99
Recall	0.97	0.97	0.93	0.97	0.97	0.97	0.97
Precision	0.92	0.9	0.9	0.9	0.94	0.93	1

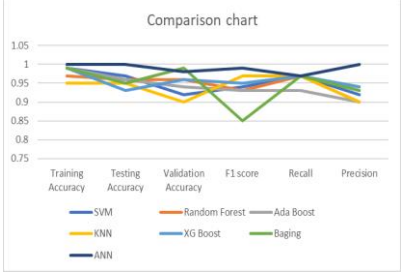


Figure 6.7: Comparative analysis of the models

7 CONCLUSION

In this work, we classified breast cancer using eight different algorithms utilizing the Wisconsin Breast Cancer dataset. To find the best approach, we collected the performance measure using 1-fold and 5-fold cross-validation. In comparison to other algorithms, the investigation showed that XGBoost had the highest AUC of 99,61 percent and the best performance statistic. We come to the conclusion that XGBoost is the most accurate algorithm for diagnosing breast cancer when using the Wisconsin Breast Cancer dataset. In the future, XGBoost can be tested against different algorithms and datasets that weren't utilized in this experiment. According to the results of the investigation, combining multidimensional data with different feature selection, classification, and dimensionality reduction techniques may provide useful tools for inference in this area. It is essential.

REFERENCES

[1] Key, T. J., Verkasalo, P. K., & Banks, E. (2001). Epidemiology of breast cancer. The lancet oncology, 2(3), 133-140.

[2] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012.

[3] Chaurasia, V., & Pal, S. (2014). Data mining techniques: to predict and resolve breast cancer survivability. International Journal of Computer Science and Mobile Computing IJCSMC, 3(1), 10-22.

[4] Djebbari, A., Liu, Z., Phan, S., & Famili, F. (2008). An ensemble machine learning approach to predict survival in breast cancer. International journal of computational biology and drug design, 1(3), 275-294.

[5] 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.

[6] Agarap, A. F. M. (2018, February). On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. In Proceedings of the 2nd international conference on machine learning and soft computing (pp. 5-9 7. V. Chaurasia, S. Pal, and B. Tiwari, "Prediction of benign and malignant breast cancer

[7] "using data mining techniques," Journal of Algorithms & Computational Technology,vol. 12, no. 2, pp. 119-126, 2018.

[8] Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. IEEE Access, 8,150360-150376.

[9] Toprak, A. (2018). Extreme learning machine (elm)-based classification of benign and malignant cells in breast cancer. Medical science monitor: international medical journal of experimental and clinical research, 24, 6537.

[10] Jacob, D. S., Viswan, R., Manju, V., PadmaSuresh, L., & Raj, S. (2018, March). Asurvey on breast cancer prediction using data miningtechniques. In 2018 Conferenceon Emerging Devices and Smart Systems (ICEDSS) (pp. 256-258). IEEE.

[11] Padhi, T., & Kumar, P. (2019, January). Breast Cancer Analysis Using WEKA. In 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 229-232). IEEE.

[12] Thomas, T., Pradhan, N., & Dhaka, V. S. (2020, February). Comparative analysis to predict breast cancer using machine learning algorithms: a survey. In 2020 International Conference on Inventive Computation Technologies (ICICT) (pp. 192-196). IEEE.

[13] Livingston, F. (2005). Implementation of Breiman's random forest machine learning algorithm. ECE591Q Machine Learning Journal Paper, 1-13.

[14] Mitchell, R., & Frank, E. (2017). Accelerating the XGBoost algorithm using GPU computing. PeerJ Computer Science, 3, e127.