Current Trends in Biotechnology and Pharmacy
Vol. 16 (4) 481 - 489, Oct 2022, ISSN 0973-8916 (Print), 2230-7303 (Online)
10.5530/ctbp.2022.4.81

Automated Detection Of Breast Cancer Using Machine Learning Algorithms: A Comparative Analysis

Venkata Srinivas Babu Oguri[1], Sudhakar Poda[2]
[1]Department of Computer Science and Engineering, Mahindra University, Hyderbad, Telangana
[2]Department of Biotechnology, Acharya Nagaarjuna University, Nagarjunanagar, Guntur-522510, India
Corresponding author: Sudhakar Poda@hotmail.com

Abstract

Breast Cancer is one of the most commonly diagnosed cancers in women. There is a 12.6% chance that a women can develop invasive breast cancer in her lifetime. All the clinically performed methods have some limitations. Automated diagnosis became an important area of cancer studies. Development of such systems require machine learning algorithms. Machine Learning has became a popular tool in the field of medical diagnosis. This study aims to provide a comparative analysis of various algorithms that can be used in the classification of Breast Cancer. The dataset used was retrieved from UCI machine learning repository and was created by University of Wisconsin. A total of 9 algorithms were employed to classify Benign and Malignant Tumors. 9 models were created using these algorithms. These models were compared by various metrics such as Accuracy, Precision, Recall Score and F1 Score.

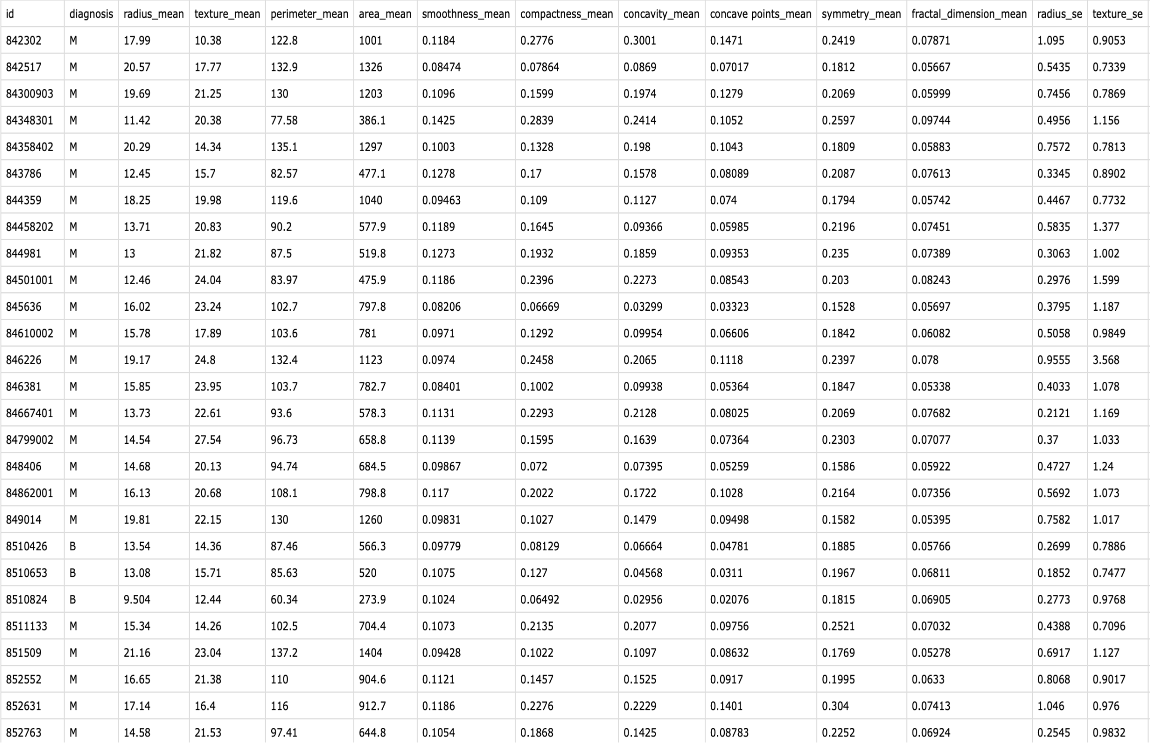
Keywords Breast Cancer, LightGBM, Jupyter notebook, Pearson Correlation, ROC curve, Accuracy, F1 Score

Introduction

Breast Cancer is one of the most commonly diagnosed cancers in women. There is a 12.6% chance that a women can develop invasive breast cancer in her lifetime(1). 95% of Breast Cancers arise from breast epithelial elements and are carcinomas. Two main types of Breast Cancer are invasive carcinomas and *in situ* carcinomas. Breast cancer cases are high in ages upto 50(2). Genetics might play an important factor as 6% of the cases are hereditary(3). Also, women diagnosed with breast Cancer have higher risk factor of developing new cancer in either the second or the treated breast. This risk factor will increase by 1% every year. Much research had been done trying to establish a relationship between specific foods and diagnosis of Breast Cancer. The only reliable relation identified was with alcohol(4). Furthermore, according to Nurses Health Society, postmenopausal women who exercised at least one hour per week are 20% less likely to be diagnosed with Breast Cancer. Clinical diagnosis of this disease is generally done by physical examination, mammography and ultrasound(5). All these methods have some limitations. Ultra sound may not detect cancer at all times. Also, analysis of mammographic image is difficult as it shows little contrast between tumors and normal tissue(6). Early diagnosis helps in delaying of the disease progress and limits treatment expenses. Automated diagnosis became an important area of cancer studies. Development of such systems require machine learning algorithms. Several studies were involved in automated cancer detection using various algorithms(7)(8). Machine Learning has become a popular tool in the field of medical diagnosis. This study aims to provide a comparative analysis of various algorithms that can be used in the classification of Breast Cancer.

Material and Methods

The dataset was retrieved from UCI machine learning repository. It has a total of 32 attributes and 569 instances. The creators of this dataset were Dr. William H. Wolberg, W. Nick Street and Olvi L. Mangasarian of University of Wisconsin. The attributes were calculated from Digitalised image of FNA of a breast mass. Fig1 shows a section of the dataset. In the diagnosis attribute, B stands for Benign while M stands for malignant. Fig2 and Fig3 depict the relation between Texture Mean, Radius Mean and Area Mean, Concavity Mean. The dataset underwent pre-processing initially. Correlation coefficients were calculated between every attribute. The correlation matrix is shown in Fig4. The correlation Coefficient has values ranging from -1 to 1. Highly correlated attributes will have values closer to 1, while uncorrelated ones will have values closer to 0. Attributes with coefficients greater than 0.9 were removed. Out of 31 features (excluding diagnosis), 10 were removed and new dataset was created. Next, this dataset was split into testing and training data. 20% was set aside for testing and 80% was used to train the model. A total of 9 algorithms were used to create 9 models. Each of these models were fit(trained) with the training data. Various parameters for accessing the performance of each of the model were also calculated. Random Forest, Decision Tree, XGBoost, LightGBM, Naive Bayes, Logistic regression, KNN, SVM and AdaBoost classifiers were the algorithms used in this study.



id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean	radius_se	texture_se
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053
842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339
84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902
844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377
844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002
84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599
845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187
84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849
846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568
846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078
84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169
84799002	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033
848406	M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24
84862001	M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073
849014	M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017
8510426	B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.7886
8510653	B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1852	0.7477
8510824	B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	0.2773	0.9768
8511133	M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	0.4388	0.7096
851509	M	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	0.6917	1.127
852552	M	16.65	21.38	110	904.6	0.1121	0.1457	0.1525	0.0917	0.1995	0.0633	0.8068	0.9017
852631	M	17.14	16.4	116	912.7	0.1186	0.2276	0.2229	0.1401	0.304	0.07413	1.046	0.976
852763	M	14.58	21.53	97.41	644.8	0.1054	0.1868	0.1425	0.08783	0.2252	0.06924	0.2545	0.9832

Fig1 Dataset

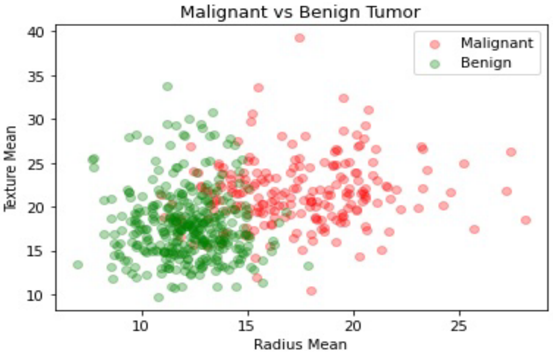


Fig 2:-Relationship between Texture and Radius Mean

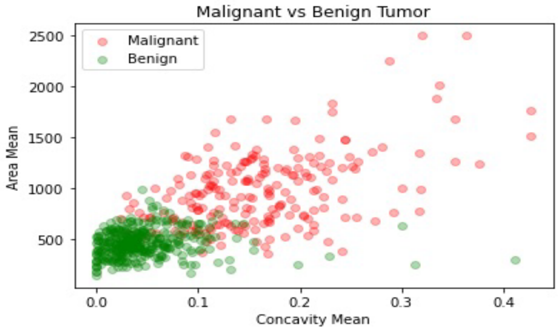


Fig 3:-Relationship between Area and Concavity Mean

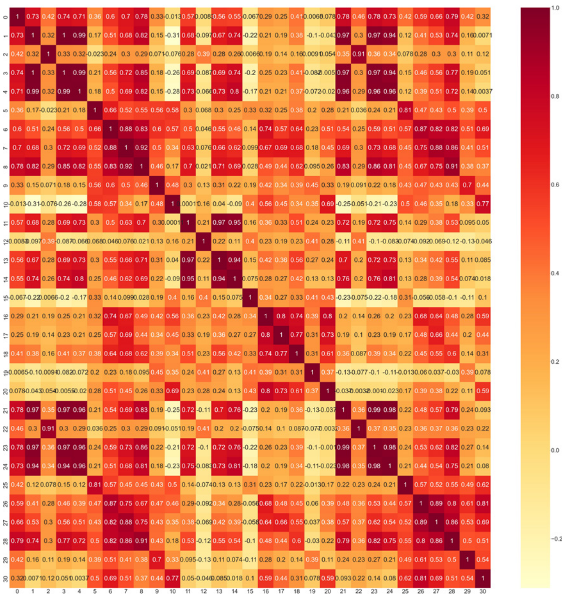


Fig 4:-Correlation Matrix

In Fig 4 colour grading was set to 'yellow red'. High correlation is shown by more red colour, while Low correlation is shown by more yellow colour. All diagonal elements have correlation value of 1 and are red since each feature is mapped to itself. Fig 5 shows the scatter plot of the new dataset.

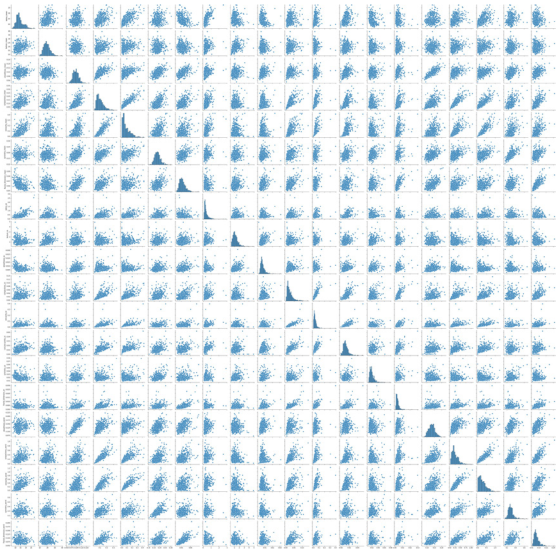


Fig 5:- Scatter Plot

Brief descriptions of various algorithms used in this study are described below.

Random Forest classifier

Random Forest contains a large number of decision trees, where each tree gives out a prediction and the one with highest number of votes will be given out as output(9). It works with the fundamental concept that a large number of Trees(models) working together will outperform any individual constituent models.

XGBoost classifier

XGBoost works under a gradient boosting framework and is also Decision Tree based. It was developed at University Of Washington in 2016. Similar to Random Forest, it is an ensemble learning algorithm. It's known for its handling of data types , distributions and the variety of hyper parameters that can be tuned(10).

Adaboost classifier

AdaBoost is one of the first boosting algorithms created. It combines multiple weak learners into a single strong learner and can be used for both regression and classification. It works by giving more preference on ones more difficult to classify instances rather than on easily handled ones(11).

Decision tree classifier

Decision Tree Classifier comes under supervised learning techniques and is mainly used for classification. Its structure represents a tree where the internal nodes represent the features, the branches specify the decision rules and the leaf nodes present the outcome(12). Fig 6 illustrates the decision tree generated on this dataset.

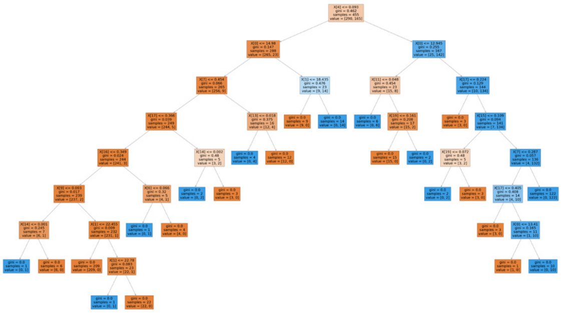


Fig 6 :-Illustration of Decision Tree

K-nearest neighbours classifier

K-Nearest Neighbours or shortly KNN is also a supervised learning technique used for classification as well as regression. It can also be categorised under Lazy Learner Algorithms as doesn't immediately learn from the training data. It tires to calculates the similarity between new and trained data and put the new case into one of the categories(13).

Support vector machine classifier

Support Vector Machine(SVM) creates a best line or boundary that can separate N-dimensional space in categories so that new data can easily be put in one of the correct class. It chooses Extreme Vectors which help in generating a Hyperplane. These Extreme Vectors are called Support Vectors(14).

LightGBM

LightBGM share many similarities with XGBoost in features such as parse optimisation, parallel training, multiple loss functions and bagging. LightBGM grows trees leaf-wise instead of level- wise(15). It uses histogram based trees instead of Decision Trees. It was developed by Microsoft.

Logistic regression

Logistic Regression is a supervised learning algorithm and is used mainly for binary classification. It uses sigmoid function to classify the data and can easily determine the most effective attributes used for classification. It provides probable values between 0 and 1 (16).

Naive bayes classifier

Naive Bayes is a classification algorithm based on Bayes Theorem. It naively assumes independence among the features and the continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution(17).

All the work was done using Jupyter Notebook, an interactive python notebook. All the algorithms used were performed using Scikit-learn machine learning library. Seaborn library was used to visualise the data.

Results and Discussion

Naive Bayes and Decision Tree Classifiers were the least performing models with an accuracy of 89.47% and 88.59% respectively. LightGBM achieved the highest accuracy of 98.24%. AdaBoost and XGBoost performed well with an accuracy of 96.49% and 97.36%. Fig 7 to Fig 15 present the confusion matrices of all algorithms used. Confusion Matrix gives us the number of True positive(TP) ,True negative(TN), False positive(FP) and False negative(FN) instances.

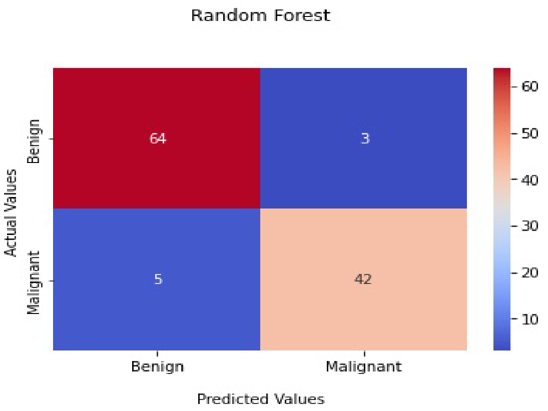


Fig 7

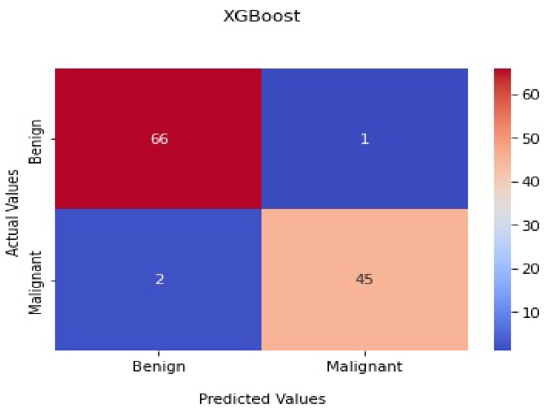


Fig 8

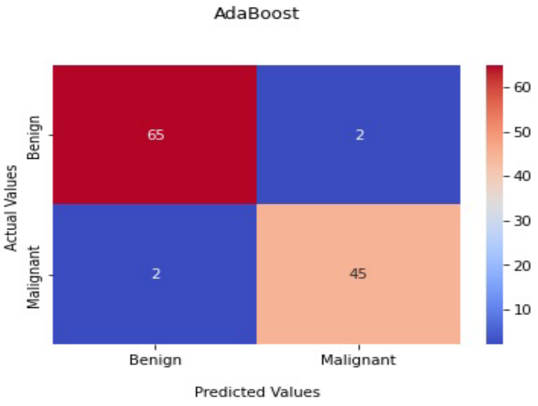


Fig 9

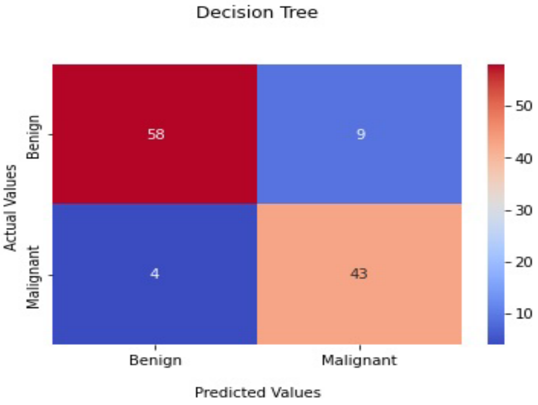


Fig 10

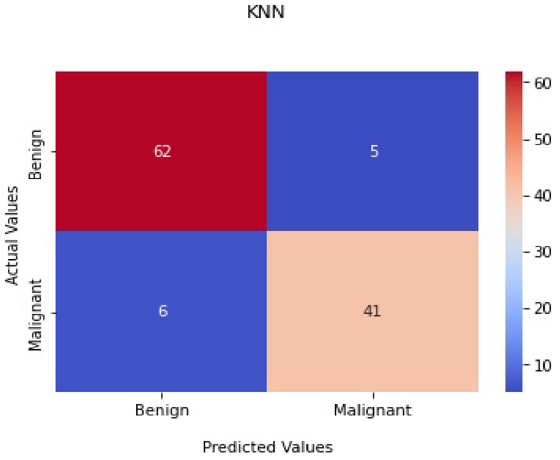


Fig 11

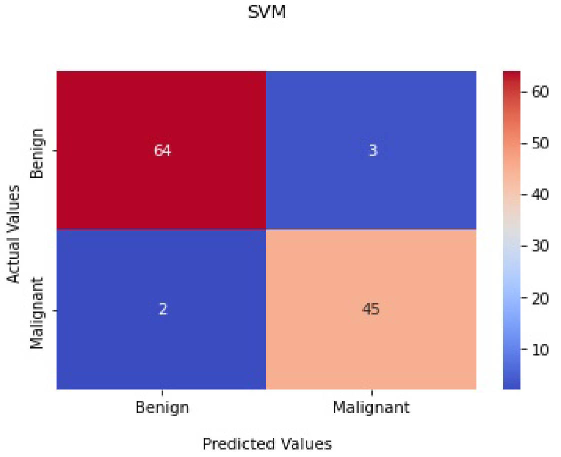


Fig 12

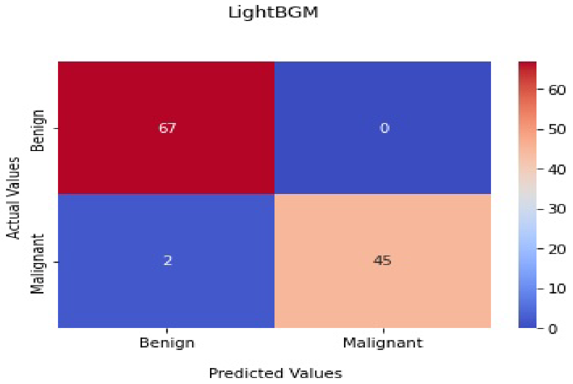


Fig 13

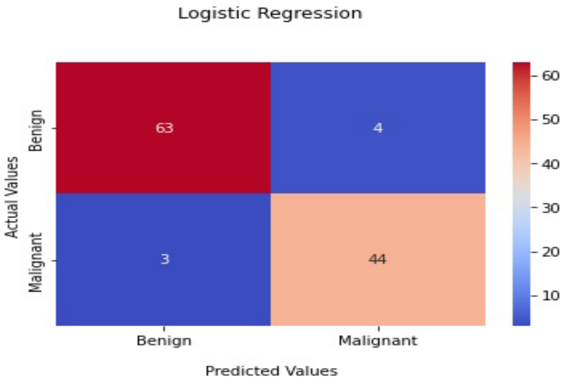


Fig 14

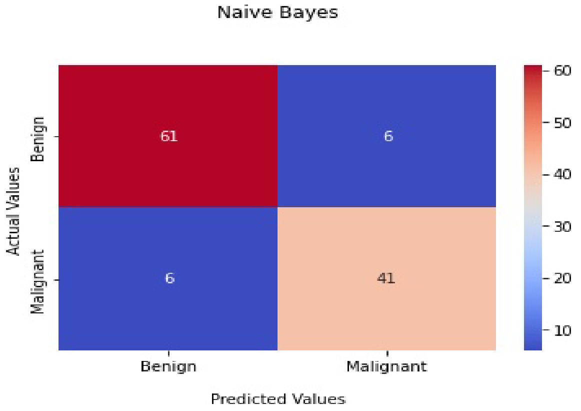


Fig 15

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

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

$$recall = \frac{TP}{TP + FN}$$

$$F1 - Score = \frac{2 \times Precision \times recall}{Precision + recall}$$

Random Forest classifier has shown 42 True Positive, 64 True Negative, 3 False Positive and 5 False Negative instances. A Total of 106 instances have been correct out of 114, hence the accuracy is 92.98%.Precision, Recall and F1 Score were also calculated for each algorithm. Their Formulas are given Above. Each of these metrics were calculated for all the nine algorithms used and their values are given in Table 1. All values were compared graphically in Fig 16.

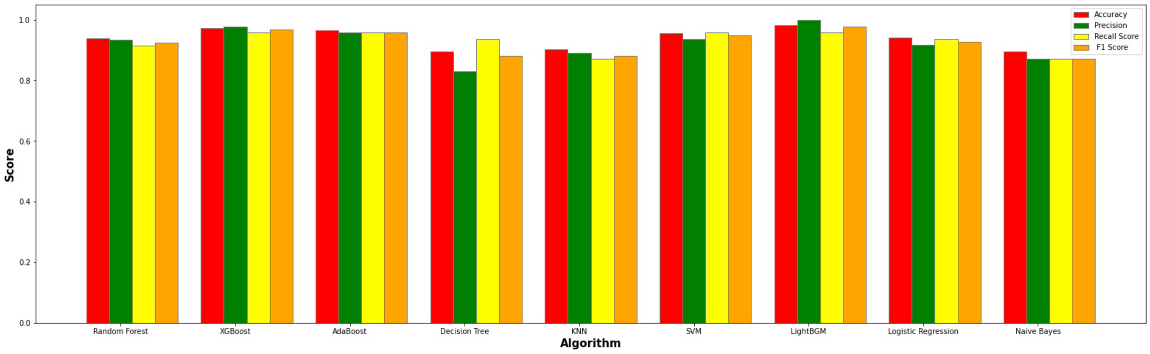


Fig 16:-Comparison of Metrics

Table 1:- Values Of Metrics

Algorithm	Accuracy	Precision	Recall Score	F1_Score
Random Forest	0.938596	0.934783	0.914894	0.924731
XGBoost	0.973684	0.978261	0.957447	0.967742
AdaBoost	0.964912	0.957447	0.957447	0.957447
Decision Tree	0.894737	0.830189	0.936170	0.880000
KNN	0.903509	0.891304	0.872340	0.881720
SVM	0.956140	0.937500	0.957447	0.947368
LightGBM	0.982456	1.000000	0.957447	0.978261
Logistic Regression	0.940659	0.916667	0.936170	0.926316
Naive Bayes	0.894737	0.872340	0.872340	0.872340

LightGBM scored the highest F1 Score of 97.82%, followed by XGBoost and AdaBoost with 96.77% and 95.74% respectively. SVM, AdaBoost, LightGBM and XGBoost achieved Recall Score of 95.74%. In all the four parameters, LightGBM performed better than the remaining algorithms. Also, boosting algorithms outperformed tree based algorithms in all parameters. ROC (Receiver Operating Characteristic) curves along with the AUC (Area Under The Curve) are presented in Fig 17-25.

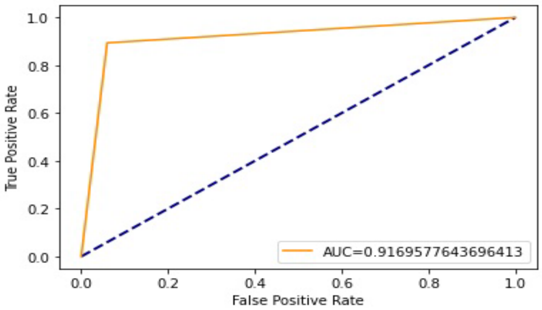


Fig 17:- Random Forest

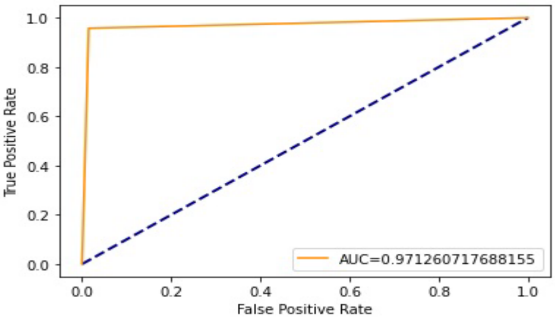


Fig 18:- XGBoost

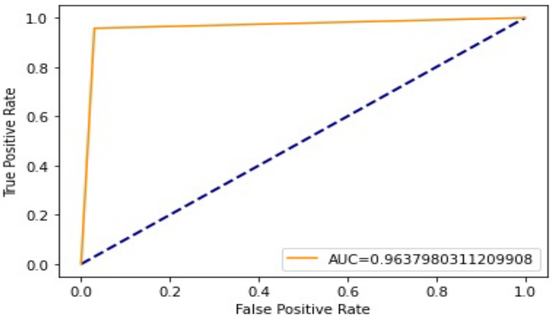


Fig 19:- AdaBoost

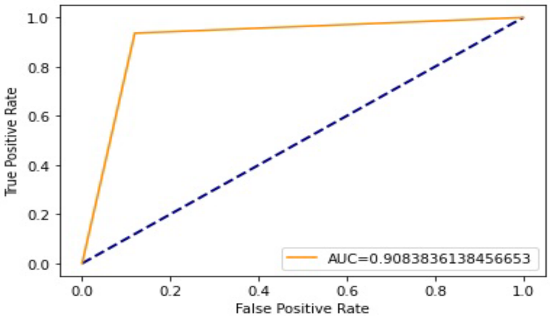


Fig 20:- Decision Tree

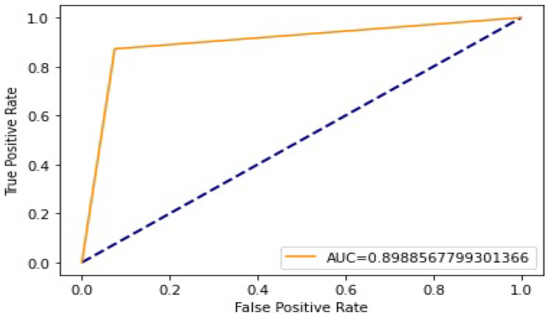


Fig 21:- KNN

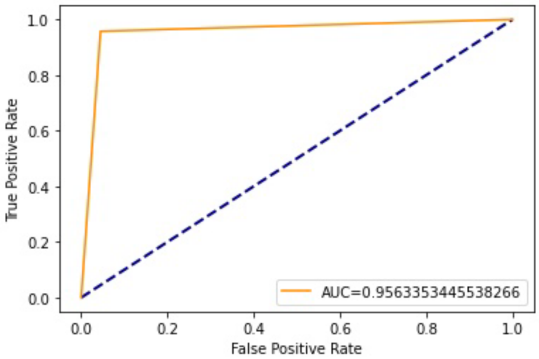


Fig 22:- SVM

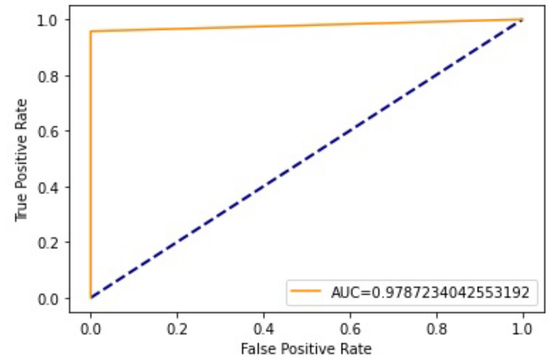


Fig 23:- LightGBM

ROC curves are used in binary classification and feature True Positive and False Positive rates on x and y axes respectively. A larger AUC is considered better. LightBGM has the highest AUC of 97.87% while Naive Bayes has the lowest of 89.13%.

Conclusion

Breast Cancer is one of the most frequently occurring cancers in the world. This study provides a comparative analysis of different machine learning algorithms used in detection of the disease. Most of the algorithms achieved an accuracy of score higher than 90%, proving that machine learning plays a crucial role in early diagnosis and automated detection of the disease. Boosting algorithms used in study (XGBoost, LightGBM and AdaBoost) performed superiorly.

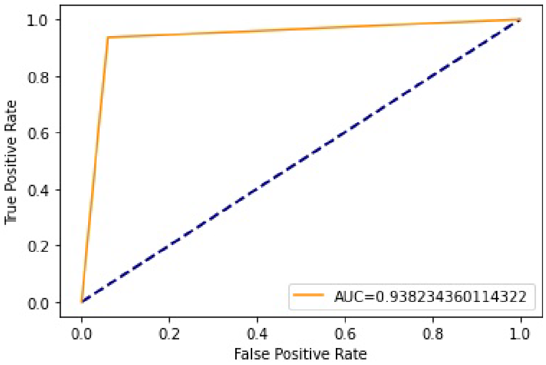


Fig 24:- Logistic Regression

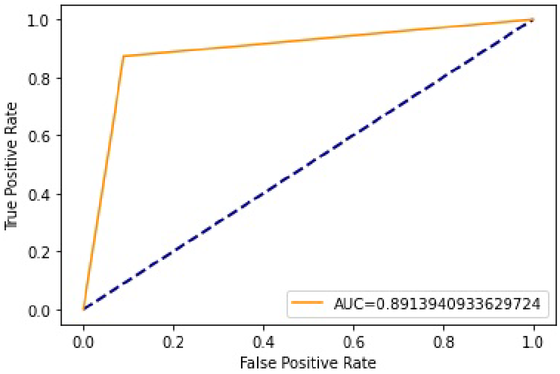


Fig 25:- Naive Bayes

References

(1) Greenlee RT, Hill-Harmon MD, Murry T, Thun M. Cancer Statistics, 2001. CA Cancer J Clin. 2001;51: 15

(2) Smith H, Kammerer-Doak D, Barbo D, Sarto G. Hormone Replacement Therapy in the Menopause: A Pro Opinion. CA—A Cancer Journal for Clinicians. 1996;46:343.

(3) Malone KE, Daling JR, Thompson JD, O'Brien CA, Francisco LV, Ostrander EA. BRCA1 mutations and breast cancer in the general population: analysis in women before age 35 years and in women before age 45 years with first-degree family history. JAMA. 1998;279:922–929.

(4) Mourouti N., Kontogianni M.D., Papavagelis C., Panagiotakos D.B. Diet

and breast cancer: A systematic review. Int. J. Food Sci. Nutr. 2015;66:1–42. doi: 10.3109/09637486.2014.950207. (PubMed) (CrossRef) (Google Scholar)

(5) Ward RC, Lourenco AP, Mainiero MB. Ultrasound-guided breast cancer cryoablation. Am J Roentgenol. 2019;213(3). doi: 10.2214/AJR.19.21329 (PubMed) (CrossRef) (Google Scholar) (Ref list))

(6) Obenauer S, Luftner-Nagel S, von Heyden D, Munzel U, Baum E, Grabbe E. Screen film vs full-field digital mammography: Image quality, detectability and characterization of lesions. Eur Radiol. 2002;12(7). doi: 10.1007/s00330-001-1269-y (PubMed) (CrossRef) (Google Scholar) (Ref list)

(7) Lughezzani G, Briganti A, Karakiewicz PI, et al. Predictive and prognostic models in radical prostatectomy candidates: A critical analysis of the literature. Eur Urol. 2010;58(5). doi: 10.1016/j.eururo.2010.07.034 (PMC free article) (PubMed) (CrossRef) (Google Scholar) (Ref list)

(8) Zhang Q, Zhao K, Song L, et al. A Novel Apoptosis-Related Gene Signature Predicts Biochemical Recurrence of Localized Prostate Cancer After Radical Prostatectomy. Front Genet. 2020;11. doi: 10.3389/fgene.2020.586376 (PMC free article) (PubMed) (CrossRef) (Google Scholar) (Ref list)

(9) Amaratunga D, Cabrera J, Lee Y-S (2008) Enriched random forests. Bioinformatics 24:2010–2014

(10) Liu, Jialin, Jinfa Wu, Siru Liu, Mengdie Li, Kunchang Hu, and Ke Li. "Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model." Plos one 16, no. 2 (2021): e0246306.

(11) Robert E. Schapire. The strength of weak learnability. Machine Learning, 5(2):197-227, 1990

(12) Banu GR. A Role of decision Tree classification data Mining Technique in Diagnosing Thyroid disease. International Journal of Computer Sciences and Engineering. 2016;4(11):111-5.

(13) D. Wettschereck and D. Thomas G, "Locally adaptive nearest neighbour algorithms", Adv. Neural Inf. Process. Syst., pp. 184-186, 1994.

(14) Song Mingjun and S. Rajasekaran, "A greedy algorithm for gene selection based on SVM and correlation", International Journal of Bioinformatics Research and Applications, vol. 6, no. 3, pp. 296-307, 2010.

(15) Song, Jiazhi, Guixia Liu, Jingqing Jiang, Ping Zhang, and Yanchun Liang. "Prediction of Protein–ATP Binding Residues Based on Ensemble of Deep Convolutional Neural Networks and LightGBM Algorithm." International Journal of Molecular Sciences 22, no. 2 (2021): 939.

(16) Y. Bi and D. R. Jeske. The efficiency of logistic regression compared to normal discriminant analysis under class-conditional classification noise. Journal of Multivariate Analysis, 101(7):1622–1637, 2010

(17) P. Langley, W. Iba and K. Thompson, "An analysis of Bayesian Classifiers.," in Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA, 1992.