

Machine learning models applied in analyzing breast cancer classification accuracy

Anuja Bokhare, Puja Jha

Symbiosis Institute of Computer Studies and Research (SICSR), A Constituent of Symbiosis International (Deemed University),
Atur Centre, Ghokhale Cross Road, Model Colony, Pune, Maharashtra, India

Article Info

Article history:

Received Jul 5, 2021

Revised Aug 14, 2021

Accepted Mar 14, 2022

Keywords:

Breast cancer

Decision tree classifier

k-nearest neighbor

Logistic regression

Naïve Bayes

Random forest

Support vector machine

ABSTRACT

There have been many attempts made to classify breast cancer data, since this classification is critical in a wide variety of applications related to the detection of anomalies, failures, and risks. In this study machine learning (ML) models are reviewed and compared. This paper presents the classification of breast cancer data using various ML models. The effectiveness of models comparatively evaluated through result using benchmark of accuracy which was not done earlier. The models considered for the study are k-nearest neighbor (kNN), decision tree classifier, support vector machine (SVM), random forest (RF), SVM kernels, logistic regression, Naïve Bayes. These classifiers were tested, analyzed and compared with each other. The classifier, decision tree, gets the highest accuracy i.e. 97.08% among all these models is termed as the best ML algorithm for the breast cancer data set.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Anuja Bokhare

Symbiosis Institute of Computer Studies and Research (SICSR), A Constituent of Symbiosis International (Deemed University)

Atur Centre, Ghokhale Cross Road, Model Colony, Pune-411016, Maharashtra State, India

Email: anuja.bokhare@gmail.com

1. INTRODUCTION

Artificial intelligence (AI) is basically a combination of engineering, computer science and other applied sciences. It has the capability to solve complex problems that would traditionally require human intelligence [1]. In this field we can say that machine learning (ML) has progressed rapidly and it can be assumed that ML can solve many real-world problems. With respect to this researcher are researching on the first presentations of AI solutions to tumor imaging and classifying that can produce a major technology change in the field of oncology [2]–[6]. ML is the part of AI which deals with scientific study of statistical model and algorithms that computer uses for doing its specific task. Current study focuses on breast cancer data classification by means of popular ML models.

Section 2 focuses on related study and its progress in classification of disease. Also discuss about various ML models. Section 3 explain the proposed model. Section 4 highlights about method used. Section 4 focus on result analysis and discussion. Section 5 is devoted for conclusion.

2. PREVIOUS BACKGROUND

Rahman and Muniyandi [7] have proposed a method to classify cancer with more accuracy and efficiency using a two-step feature selection (FS) technique combined with the 15-neuron neural network. In their research they have utilized dataset from the Wisconsin diagnostic breast cancer, in which the FS

technique was used to reduce the number of attributes while the 15-neuron neural network classified the cancer. The FS technique helped reduce the data dimensionality there by increasing the accuracy of the classification to 99.4% which is significantly higher compared to the existing papers. Tanha *et al.* [8] have tried to identify the various relationships among the various factors of different breast cancer groups. Data mining techniques and classification algorithms on dataset containing 624 patients from Iran. Using binary tree and rule-based classification, a model is developed using patterns discovered during training and testing, while establishing the significant relationships among the different prognostic indices.

A new method for classifying cancer has been proposed [9]. Naïve Bayes classification technique is considered to be very accurate due to the assumption that parameters and predictors are conditionally independent. But this also led to a loss of accuracy. The authors have proposed a method which uses hoeffding tree method for normal classification and then Naïve Bayes classification for reducing Data dimensionality. This separation technique proved to be quite effective and was able to identify input as benign, non-benign and normal with significant accuracy. How the use of anthropometric data and collected routine blood analysis parameters can estimate the prospective of having Breast cancer have been discussed in the study [10]. By using artificial neural network (ANN) and Naïve Bayes classification on data and using routinely acquired and controlled parameters, the accuracy of diagnostics is notably high. Therefore, using either ANN or Naïve Bayes technique it is possible to detect breast cancer early.

The various ML techniques and study their levels of accuracy in FS have been collated in the study [11]. The various Techniques were used on the dataset from the Wisconsin diagnostic breast cancer and appraise the accuracy of runtime results, classification test and standardized evaluation. The six ML techniques discussed in the paper include classification and regression techniques. The authors have attempted to fluctuate the various datasets and apply the various techniques to the sub-datasets. 80% of the dataset was used in training and 20% was used in testing while accuracy was acquired from voting classifier. All ML techniques showed more than 90% accuracy at various levels of classification, especially on subsets of data. Darabi *et al.* [12] have discussed how using deep learning techniques can lead to, various methods for automatic detection of breast cancer using histopathological images. Authors have proposed a well-founded model based on deep transfer learning in which the deep convolutional neural network (DCNN) is pre-conditioned using a well-constructed and notable assortment of ImageNet dataset and use data augmentation in order to uncover malignant or benign sample tissues. This includes both binary and multiclass classification and by examining the model using optimized hyper-parameters. They have developed a significantly effective transfer learning architecture with pre-conditioned DenseNet121 and ResNet50 models combined with a fully-connected classifier. Authors have used hybrid minimum redundancy maximum relevance (mRMR) and random subset feature selection (RSFS) algorithms for feature selection with k-nearest neighbor (kNN) and support vector machines (SVMs) algorithms as classifier where accuracy of 77.41% and 73.07%, and sensitivity of 98% and 72.72%, respectively is achieved [13].

Yan *et al.* [14] propound a technique using histopathological images to identify breast cancer cells which uses new hybrid convolutional and recurrent deep neural network. Proposed method uses recurrent neural network (RNN) extractor, which is second to the convolutional neural network (CNN) extractor, while considering the short- and long-term spatial correlations between the pathological image patches. Authors have also made the dataset they worked to the public, and is one of the largest datasets available to the public and consists of 3771 breast cancer histopathological images, which are quite inclusive, diverse and covering various subclasses. In [15] Deep learning allows multi task learning that reduces the losses by applying different algorithms. Deep learning also allows multimodal learning that is process to integrate different types of data. In this process AI helps in combination of different types of data from different data source. By using deep learning technique, we can do quantitative examination of number of patient's types of techniques used. We can also do qualitative assessments.

If we consider the field of oncology following challenges are observed during the study.

- Minimum no of datasets available: There are few datasets available with this few dataset the performance and accuracy of the model can't be determined. This non availability of dataset is the biggest barrier that is preventing oncologist to use AI in their treatment.
- Applying nonlinear techniques: If we apply nonlinear techniques of ML algorithm it may be applied to nonlinear phenomenon in oncology but it may degrade the utility of and performance of the dataset
- Unbiased training: Due to limited datasets it is difficult to conduct unbiased training of ML algorithm
- Lack of quality datasets: Due to lack of voluminous quality datasets it is difficult to determine the accuracy
- Lack of security: The worse challenge faced by AI in oncology is the privacy and security
- Lack of accuracy: The blackbox nature of AI model makes it difficult to interpret the accurate result which might cause unintended harm

- Larger expectations: There have been a misconception of AI capabilities. It is expected AI to perform more efficient task more than its capabilities.

AI is used to extract phonemics characteristics from medical imaging data that is called radionics that has very significant meaning and is the most popular area of research in AI. Radionics started with engineering features and ML algorithms. ML builds a mathematical data on the sample data available and make predictions to solve the complex tasks. Following is the brief overview of the ML models which are used in the classification process during the study.

Logistic regression is very popular supervised ML algorithm. It is used for predicting categorical dependent variable. The output is either 0 or 1 or yes or no [16]. It gives the probabilistic values that lies between 0 and 1. Logistic regression is much similar in working with linear regression but its working differs in various aspect. Linear Regression is used for solving regression problem and logistic regression helps with classification problem in logistic regression. In linear regression the model is fitted with straight line and in logistic regression the model is fitted with s shape curve [17], [18].

The kNN is the simplest and easy supervised ML algorithm that is easy to implement. kNN are used to find k nearest neighbors of data point in dataset. kNN helps to classify a new datapoint to the nearest class. kNN finds the similarity between new cases and the available cases and put the new classes to the category that satisfies the similarity of nearest class. These algorithms are used in many applications like texture selection [19], clustering [20], classification [21], pattern recognition [22]–[24].

Decision tree is a supervised ML that helps in both classification and regression trees (CART) that is classification and regression tree. It is a tree like structure. Tree starts with a root node and each node represents a, each leaf node represents a class label and branches represents concurrence of features that lead to class [25]. High speed and efficiency are two advantages of decision tree. One can take decision on the characteristics or the features of the dataset. It is a pictorial representation of the given conditions and gives us the best solution [25]. The decision is performed on the basis of characteristics of the dataset. [26]. Decision tree is a most widespread algorithm for estimation, prediction and classifying patterns [27].

SVM is the most popular supervised ML algorithm. It is a type of algorithm that use to classify two group classification problem. SVM is used in image classification. Goal of SVM algorithm is to create the best decision. SVM is used for both classification and regression [28]–[30]. The best decision boundary is called hyperplane SVM chooses the extreme points that are called support vectors. Complexity depends on support vectors and the accuracy of SVM is increased by reducing the dimensionality of the dataset [30]–[33].

SVM algorithm uses a group of mathematical functions. This function is to require data as input and transform it to the desired form. The SVM kernel functions returns the scalar product return between the two datapoints that has maximum space. SVM kernels that can be used for univariate and multivariate time series audits [34]. A multi kernel SVM algorithm is proposed to classify accuracy of brain tumor [35].

Naïve Bayes a based-on Bayes theorem. It is supervised ML algorithm. It helps to classify problem. Prediction of class from unknown dataset is performed using Bayes theorem. It is mainly used for classification including high dimensional training set. It is the most effective algorithm that helps in constructing fast ML models and makes quick predictions. It is a probabilistic theorem that gives output in probability [36]. It is basically a classification that is probability based its prediction does not have surety of the output that is output is not much reliable. Sentimental analysis, classifying articles and spam filtration are few popular applications of Naïve Bayes algorithm [37], [38].

Random forest (RF) algorithm forms a family of classification algorithm that is a collection of decision tree. The main component of RF is a binary tree that is made using recursive partitioning (RPART). RF helps in both classification and regression. It can only classify the data into two classes from the root node to its two offspring nodes so that similarity and consistency is maintained. RF is often a group of hundreds to thousands of trees, where each tree is developed using a bootstrap example of the original data [39]. Random forest trees are different from classification and regression as they are developed non deterministically and it uses a two-stage random process. This algorithm is used in spatial prediction [40], to assess groundwater potential [41], and Gene Ontology [42]–[44].

3. PROPOSED RESEARCH MODEL

The breast cancer database is a publicly available dataset from the University of California Irvine Machine Learning Repository. It gives information on cancerous features such as size, density and the study made here is focused primarily on comparing the classification results of tumors into malignant and benign with accuracy. Three classification algorithms are performed in this comparative study which are logistic regression, RF and decision tree. For each cell nucleus real value features which are 10 are considered. For attribute class numbers 0 and 1 were set to signify benign and malignant tumours respectively. Figure 1 explains the classification model used during the comparative study.

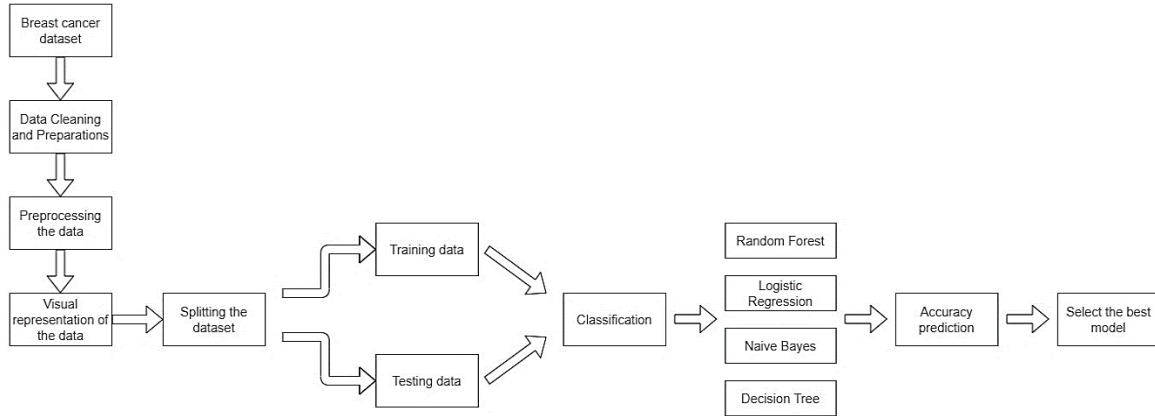


Figure 1. Classification model

4. METHOD AND EXPERIMENT DETAIL

To implement all ML model wisconsin breast cancer dataset provided by UCI Machine Learning Repository is used during the study. It has 699 examples, 2 classes i.e. malignant and benign, and 10 attributes i.e., concavity of the tumour (severity of concave portions of the contour), radius of the tumour (mean of distances from centre to points on the perimeter), texture of the tumour (standard deviation of grey-scale values), concave points (number of concave portions of the contour), symmetry, fractal dimension ("coastline approximation" - 1), Smoothness of the tumour (local variation in radius lengths), compactness of the tumour (perimeter² / area - 1.0), perimeter of the tumour, area of the tumour Figure 2 shows the snapshot of the breast cancer dataset.

id	diagnosis	radius_m	texture_n	perimeter	area_mea	smoothne	compactn	concavity	concave_p	symmetry	fractal_dii	radius_se	texture_si	perimeter	area_se	smoothne	compactn	concavity	concave_p
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	8.589	153.4	0.006399	0.04904	0.05373	0.01587
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	3.398	74.08	0.005225	0.01308	0.0186	0.0134
84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832	0.02058
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	3.445	27.23	0.00911	0.07458	0.05661	0.01867
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	5.438	94.44	0.01149	0.02461	0.05688	0.01885
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	2.217	27.19	0.00751	0.03345	0.03672	0.01137
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	3.18	53.91	0.004314	0.01382	0.02254	0.01039
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	3.856	50.96	0.008805	0.03029	0.02488	0.01448
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	2.406	24.32	0.005731	0.03502	0.03553	0.01226
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	2.039	23.94	0.007149	0.07217	0.07743	0.01432
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	2.466	40.51	0.004029	0.009269	0.01101	0.007591
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	3.564	54.16	0.005771	0.04061	0.02791	0.01282
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	11.07	116.2	0.003139	0.08297	0.0889	0.0409
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	2.903	36.58	0.009769	0.03126	0.05051	0.01992
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	2.061	19.21	0.006429	0.05936	0.05501	0.01628
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	2.879	32.55	0.005607	0.0424	0.04741	0.0109
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	3.195	45.4	0.005718	0.01162	0.01998	0.01109
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	3.854	54.18	0.007026	0.02501	0.03188	0.01297
849014	M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.006494	0.01893	0.03391	0.01521
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	2.058	23.56	0.008462	0.0146	0.02387	0.01315
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	1.383	14.67	0.004097	0.01898	0.01698	0.00649
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	1.909	15.7	0.009606	0.01432	0.01985	0.01421
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	3.384	44.91	0.006789	0.05328	0.06446	0.02252

Figure 2. Breast cancer data

Algorithm steps

1. The first step is importing the necessary libraries which would be used for building the model such as NumPy, pandas, matplotlib, seaborn.
2. After that, loaded the dataset to integrated development environment.
3. Checked if there are any missing or null values.
4. Found unique values in 'diagnosis' column and replaced 'M' with 1 and 'B' with 0 in column 'diagnosis'
5. Created visualizations of diagnosis for better understanding of the data.
6. Removed highly correlated features and reduced the data frame to 24 columns.
7. Imported sklearn library for building the model and selected features and target for X and Y axis.
8. Next step is Standard Scaling where we are getting the training and test data and scaling the data for further analysis and divided the data in testing and training set in the ratio of 20:80.

9. For classification models, imported all the required libraries and used Linear regression, Random Forest, Naïve Bayes and decision tree for prediction.
10. Compared the accuracy of the four models and found the significant models for predicting the accurate analysis.

5. RESULT ANALYSIS AND DISCUSSION

This section provides detail about the experiments and results. In current study ML models are used for classification of the breast cancer whether it is malignant or benign. Table 1 shows predictive analytics that is confusion matrix. Along with this more detail analysis of accuracy, precision, recall, F1-score and support is derived and presented for comparison between different ML models Figures 3(a) to 3(g) describe the confusion matrix results obtained for kNN, decision tree, RF, SVM, Naïve Bayes, logistic regression and kernel SVM ML algorithm respectively.

Table 1. Classification result of different ML models

ML models	Performance metrics				
	Accuracy	Precision	Recall	F1-score	Support
kNN	95.68%	95.02%	97.02%	76.00%	60%
Decision tree	97.08%	97.28%	97.45%	77.52%	62%
RF	95.6%	96.24%	97.01%	77.01%	61%
SVM	95.6%	95.24%	97.00%	76.01%	60%
Naïve Bayes	94.89%	91.28%	100%	70.00%	58%
Logistic regression	95.6%	96.00%	96.54%	77.34%	61.3%
Kernel SVM	95.6%	94.01%	98.02%	75.42%	58%

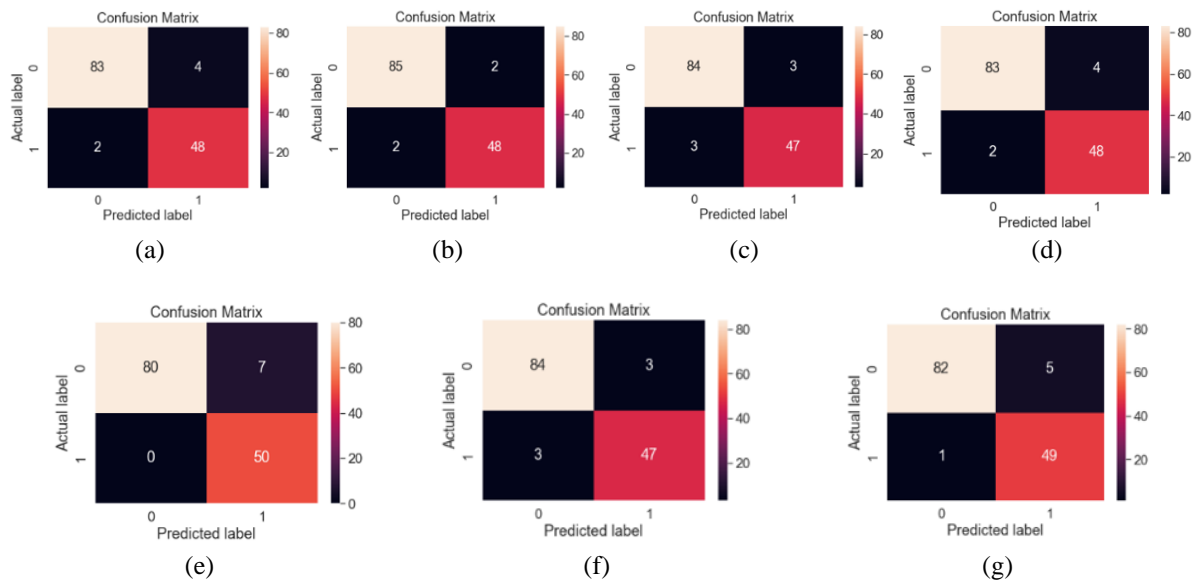


Figure 3. Confusion matrix for (a) kNN, (b) decision tree, (c) RF, (d) SVM, (e) Naïve Bayes, (f) logistic regression, and (g) Kernel SVM

Figures 4(a) to 4(g) describe the visualization about the receiver operating characteristic curve (ROC) curve obtained for kNN, decision tree, RF, SVM, Naïve Bayes, logistic regression and kernel SVM ML algorithm respectively. The investigation provided shows interesting results. The superlative classifier is the decision tree classifier. Its complete performance comes out to be the uppermost i.e. 97.08% than other algorithms. Decision tree gives high precision i.e. 97.28% which relates to the low false positive rate. Recall shows the proportion of correctly predicted positive to all observation in actual class yes, we got 97.45% recall which is good as it's above 0.5. F1-score is 77.52% for decision tree and support is 62% which is overall good as compared to other ML models. Figure 5 shows describes high accuracy gained with decision tree classifier algorithm through comparison on accuracy parameter.

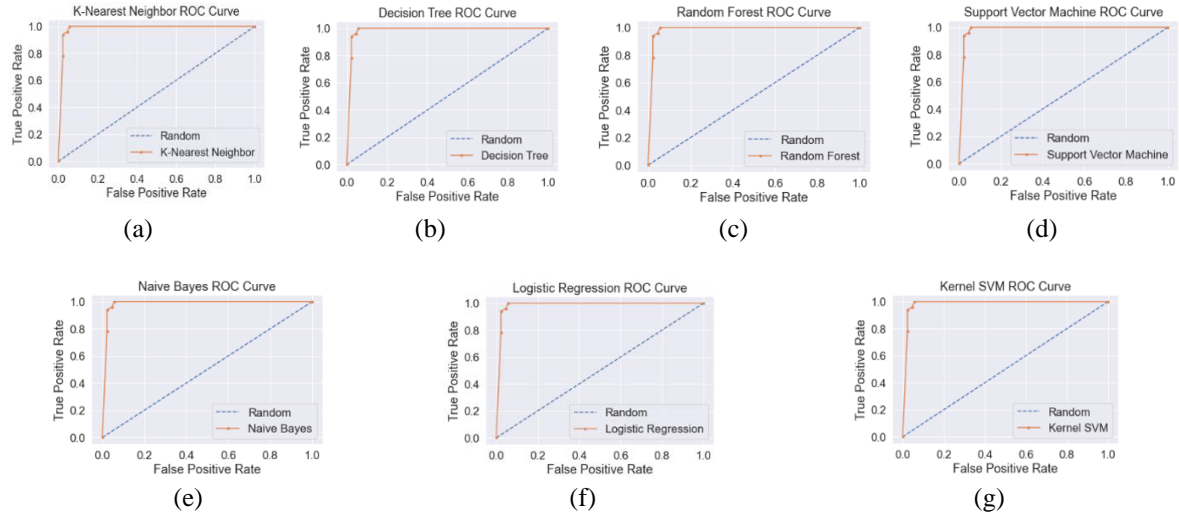


Figure 4. ROC curve for (a) kNN, (b) decision tree, (c) RF, (d) SVM, (e) Naïve Bayes, (f) logistic regression, and (g) Kernel SVM

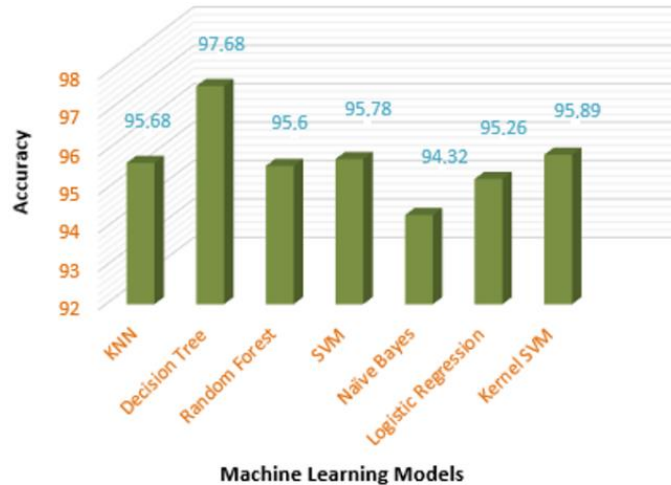


Figure 5. Accuracy comparison of different ML models

6. CONCLUSION

Current study focuses on ML models for classification of breast cancer dataset. The models considered for the study are kNN, decision tree classifier, SVM, RF, SVM kernels, logistic regression, Naïve Bayes. A comparative study of all the ML model is presented for Breast Cancer Dataset. In order to measure the performance various performance metrics are being used during the result and analysis i.e., accuracy, precision, recall F1-score and support. The obtained results show that decision tree classifier gives good accuracy 97.08% w.r.t. other ML models. Decision tree gained the knowledge about that data at the time of training itself, due to this feature decision tree is the best classifier for predicting breast cancer dataset. Accuracy obtained by applying all algorithm except decision tree is near about 95%.

REFERENCES




- [1] W. L. Bi *et al.*, “Artificial intelligence in cancer imaging: clinical challenges and applications,” *CA: A Cancer Journal for Clinicians*, vol. 69, no. 2, pp. 127–157, 2019, doi: 10.3322/caac.21552.
- [2] O. Times, “How artificial intelligence is changing oncology,” *Oncology Times*, vol. 40, no. 21, pp. 24,30-31, 2018, doi: 10.1097/01.cot.0000549549.58401.8d.
- [3] C. Luchini, A. Pea, and A. Scarpa, “Artificial intelligence in oncology: current applications and future perspectives,” *British Journal of Cancer*, vol. 126, no. 1, pp. 4–9, 2022, doi: 10.1038/s41416-021-01633-1.
- [4] R. F. Thompson *et al.*, “Artificial intelligence in radiation oncology: a specialty-wide disruptive transformation?,” *Radiotherapy and Oncology*, vol. 129, no. 3, pp. 421–426, 2018, doi: 10.1016/j.radonc.2018.05.030.

- [5] F. Azuaje, "Artificial intelligence for precision oncology: beyond patient stratification," *npj Precision Oncology*, vol. 3, no. 1, p. 6, 2019, doi: 10.1038/s41698-019-0078-1.
- [6] H. Shimizu and K. I. Nakayama, "Artificial intelligence in oncology," *Cancer Science*, vol. 111, no. 5, pp. 1452–1460, 2020, doi: 10.1111/cas.14377.
- [7] M. A. Rahman and R. C. Muniyandi, "An enhancement in cancer classification accuracy using a two-step feature selection method based on artificial neural networks with 15 neurons," *Symmetry*, vol. 12, no. 2, p. 271, 2020, doi: 10.3390/sym12020271.
- [8] J. Tanha, H. Salarabadi, M. Aznab, A. Farahi, and M. Zoberi, "Relationship among prognostic indices of breast cancer using classification techniques," *Informatics in Medicine Unlocked*, vol. 18, p. 100265, 2020, doi: 10.1016/j.imu.2019.100265.
- [9] R. A. Ibrahim Alhayali, M. A. Ahmed, Y. M. Mohialden, and A. H. Ali, "Efficient method for breast cancer classification based on ensemble hoeffding tree and Naïve Bayes," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 2, pp. 1074–1080, 2020, doi: 10.11591/ijeecs.v18.i2.pp1074-1080.
- [10] A. Yasar and M. M. Saritas, "Performance analysis of ANN and Naive Bayes classification algorithm for data classification," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 2, pp. 88–91, 2019, doi: 10.18201/ijisae.2019252786.
- [11] V. Chaurasia and S. Pal, "Applications of machine learning techniques to predict diagnostic breast cancer," *SN Computer Science*, vol. 1, no. 5, p. 270, 2020, doi: 10.1007/s42979-020-00296-8.
- [12] Y. Yari, T. V. Nguyen, and H. T. Nguyen, "Deep learning applied for histological diagnosis of breast cancer," *IEEE Access*, vol. 8, pp. 162432–162448, 2020, doi: 10.1109/ACCESS.2020.3021557.
- [13] N. Darabi, A. Rezaei, and S. S. F. Hamidpour, "Breast cancer detection using RSFS-based feature selection algorithms in thermal images," *Biomedical Engineering - Applications, Basis and Communications*, vol. 33, no. 3, p. 2150020, 2021, doi: 10.4015/S1016237221500204.
- [14] R. Yan *et al.*, "Breast cancer histopathological image classification using a hybrid deep neural network," *Methods*, vol. 173, pp. 52–60, 2020, doi: 10.1016/j.ymeth.2019.06.014.
- [15] L. Boldrini, J. E. Bibault, C. Masciocchi, Y. Shen, and M. I. Bittner, "Deep learning: a review for the radiation oncologist," *Frontiers in Oncology*, vol. 9, p. 997, 2019, doi: 10.3389/fonc.2019.00977.
- [16] S. Xia, Z. Xiong, Y. Luo, L. Dong, and G. Zhang, "Location difference of multiple distances based k-nearest neighbors algorithm," *Knowledge-Based Systems*, vol. 90, pp. 99–110, 2015, doi: 10.1016/j.knsys.2015.09.028.
- [17] S. Nusinovič *et al.*, "Logistic regression was as good as machine learning for predicting major chronic diseases," *Journal of Clinical Epidemiology*, vol. 122, pp. 56–69, 2020, doi: 10.1016/j.jclinepi.2020.03.002.
- [18] A. Ahmed, A. Jalal, and K. Kim, "A novel statistical method for scene classification based on multi-object categorization and logistic regression," *Sensors (Switzerland)*, vol. 20, no. 14, pp. 1–20, 2020, doi: 10.3390/s20143871.
- [19] J. Kuha and C. Mills, "On group comparisons with logistic regression models," *Sociological Methods and Research*, vol. 49, no. 2, pp. 498–525, 2020, doi: 10.1177/0049124117747306.
- [20] G. Poli *et al.*, "Solar flare detection system based on tolerance near sets in a GPU-CUDA framework," *Knowledge-Based Systems*, vol. 70, pp. 345–360, 2014, doi: 10.1016/j.knsys.2014.07.012.
- [21] S. Y. Xia, Z. Y. Xiong, Y. He, K. Li, L. M. Dong, and M. Zhang, "Relative density-based classification noise detection," *Optik*, vol. 125, no. 22, pp. 6829–6834, 2014, doi: 10.1016/j.ijleo.2014.08.091.
- [22] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Systems*, vol. 24, no. 7, pp. 1024–1032, 2011, doi: 10.1016/j.knsys.2011.04.014.
- [23] A. F. Seddik and D. M. Shawky, "A low-cost screening method for the detection of the carotid artery diseases," *Knowledge-Based Systems*, vol. 52, pp. 236–245, 2013, doi: 10.1016/j.knsys.2013.08.007.
- [24] M. Gohari and A. M. Eydi, "Modelling of shaft unbalance: modelling a multi discs rotor using k-nearest neighbor and decision tree algorithms," *Measurement: Journal of the International Measurement Confederation*, vol. 151, p. 107253, 2020, doi: 10.1016/j.measurement.2019.107253.
- [25] J. Gou, Y. Zhan, Y. Rao, X. Shen, X. Wang, and W. He, "Improved pseudo nearest neighbor classification," *Knowledge-Based Systems*, vol. 70, pp. 361–375, 2014, doi: 10.1016/j.knsys.2014.07.020.
- [26] M. Batra and R. Agrawal, "Comparative analysis of decision tree algorithms," *Advances in Intelligent Systems and Computing*, vol. 652, pp. 31–36, 2018, doi: 10.1007/978-981-10-6747-1_4.
- [27] K. Jacek and J. Cwik, *Statistical learning systems (in Polish: Statystyczne systemy uczące sie)*. Akademicka Oficyna Wydawnicza EXIT, 2008.
- [28] V. Vapnik, *Estimation of dependences based on empirical data*. New York, NY: Springer New York, 2006.
- [29] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/bf00994018.
- [30] D. M. J. Tax, D. de Ridder, and R. P. W. Duin, "Support vector classifiers: a first look," in *Proceedings ASCI*, 1997, vol. 97, no. im, pp. 253–258.
- [31] J. A. Gualtieri, "The support vector machine (SVM) algorithm for supervised classification of hyperspectral remote sensing data," *Kernel Methods for Remote Sensing Data Analysis*, vol. 3, pp. 51–83, 2009, doi: 10.1002/9780470748992.ch3.
- [32] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002, doi: 10.1023/A:1012487302797.
- [33] G. Anthony and H. Ruther, "Comparison of feature selection techniques for SVM classification," *10th Intl. Symposium on Physical Measurements and Spectral Signatures in Remote Sensing*, vol. 36, pp. 258–263, 2007.
- [34] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by SVM," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 5, pp. 2297–2307, May 2010, doi: 10.1109/TGRS.2009.2039484.
- [35] M. Achirul Nanda, K. Boro Seminar, D. Nandika, and A. Maddu, "A comparison study of kernel functions in the support vector machine and its application for termite detection," *Information*, vol. 9, no. 1, p. 5, Jan. 2018, doi: 10.3390/info9010005.
- [36] S. Rüping, "SVM kernels for time series analysis," *Universitätsbibliothek Dortmund*, p. 8, 2001.
- [37] S. Krishnakumar and K. Manivannan, "Effective segmentation and classification of brain tumor using rough K means algorithm and multi kernel SVM in MR images," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 6, pp. 6751–6760, 2021, doi: 10.1007/s12652-020-02300-8.
- [38] M. A. Awal, M. Masud, M. S. Hossain, A. A. M. Bulbul, S. M. H. Mahmud, and A. K. Bairagi, "A novel Bayesian optimization-based machine learning framework for COVID-19 detection from inpatient facility data," *IEEE Access*, vol. 9, pp. 10263–10281, 2021, doi: 10.1109/ACCESS.2021.3050852.
- [39] H. Zhang, N. Cheng, Y. Zhang, and Z. Li, "Label flipping attacks against Naive Bayes on spam filtering systems," *Applied Intelligence*, vol. 51, no. 7, pp. 4503–4514, 2021, doi: 10.1007/s10489-020-02086-4.




- [40] M. Abbas, K. Ali Memon, and A. Aleem Jamali, "Multinomial Naive Bayes classification model for sentiment analysis," *IJCSNS International Journal of Computer Science and Network Security*, vol. 19, no. 3, p. 62, 2019.
- [41] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata Journal*, vol. 20, no. 1, pp. 3–29, 2020, doi: 10.1177/1536867X20909688.
- [42] T. Hengl, M. Nussbaum, M. N. Wright, G. B. M. Heuvelink, and B. Gräler, "Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables," *PeerJ*, vol. 2018, no. 6, p. e5518, 2018, doi: 10.7717/peerj.5518.
- [43] S. A. Naghibi, K. Ahmadi, and A. Daneshi, "Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping," *Water Resources Management*, vol. 31, no. 9, pp. 2761–2775, 2017, doi: 10.1007/s11269-017-1660-3.
- [44] F. Fabris, A. Doherty, D. Palmer, J. P. De Magalhaes, and A. A. Freitas, "A new approach for interpreting Random Forest models and its application to the biology of ageing," *Bioinformatics*, vol. 34, no. 14, pp. 2449–2456, 2018, doi: 10.1093/bioinformatics/bty087.

BIOGRAPHIES OF AUTHORS



Dr. Anuja Bokhare    She is working as Assistant Professor in the department of Computer Science at Symbiosis Institute of Computer Studies and Research, Pune, Maharashtra India. She received M.Phil. (Computer Science) at Y.C.M.O.U, Nasik, India. She has 20 years of experience in the field of academics. Her research interest includes applications of fuzzy logic, neural network, software engineering. She had published 22 research papers in international journal and conferences along with one book in her account. She can be contacted at email: anuja.bokhare@gmail.com.



Puja Jha    has a strong analytical and quantitative skills. Ability to analyze patterns and trends in large data sets. Knowledge of data mining and data warehouse. Pursuing Post graduate degree in Information Technology at Symbiosis Institute of Computer Studies and Research, Pune, Maharashtra, India. She has work experience of 3 years in the field of data analytics. She can be contacted at email: puj1941050@sicsr.ac.in.