Int J Syst Assur Eng Manag (January 2024) 15(1):188–197
https://doi.org/10.1007/s13198-022-01696-0

ORIGINAL ARTICLE

Early predictive model for breast cancer classification using blended ensemble learning

T. R. Mahesh[1] · V. Vinoth Kumar[1] · V. Vivek[1] · K. M. Karthick Raghunath[2] · G. Sindhu Madhuri[1]

Received: 3 December 2021 / Revised: 24 February 2022 / Accepted: 4 June 2022 / Published online: 26 June 2022

Abstract Breast cancer is one of the most common cancers among women's worldwide, and it is a fact that most of the cases are discovered late. Several researchers have examined the prediction of breast cancer. Breast cancer poses a significant hazard to women. The deficiency of reliable predictive models really makes it challenging for clinicians to devise a treatment strategy that will help patients live longer. An automatic illness detection system assists medical personnel in diagnosing disease and provides a reliable, efficient and quick reaction while also lowering the danger of death. A Blended ensemble learning, which is an innovative approach, has been utilized for the classification of breast cancer and this model performs effectively for the base classifier in the prediction analysis. The performance of five machine learning techniques, namely support vector machine, K-nearest neighbors, decision tree Classifier, random forests, and logistic regression, are used as base learners in blended ensemble model. All the incorporated base learners (individually) and the final outcome of the Ensemble Learning are being compared in this study against several performance metrics namely accuracy, recall, precision and f1-score for the early prediction of Breast Cancer. There is a 98.14 percent noticeable improvement with the Ensemble Learning model compared to the basic learners.

Keywords Machine learning · Breast cancer · Accuracy · Prediction · Recall · Detection system · Diagnosis · Precision

1 Introduction

Breast cancer is a pretty common cancer amidst women all over the world. In 2016, 246,660 new records of breast cancer were predicted to be recognized in women in the United States, with 40,450 people expected to die. Breast cancer's progression and prognosis piqued my interest. As a sample set, the UCI Wisconsin ML Repository Breast Cancer Dataset attracted a high number of patients with multivariate features. The proper diagnosis of certain crucial information is a remarkable issue in the area of bioinformatics or medical research (Park and Han 2018). In the area of medicine, disease diagnosis is a demanding and challenging task. Many diagnostic institutions, hospitals, research centers, as well as numerous websites have a vast amount of medical diagnostic data. However, it is barely essential to categorize them to automate and speed up disease diagnosis. According to the American Cancer Society (Breast Cancer 2018), breast cancer affects more women than any other cancer. In 2017, approximately 252,710 women were diagnosed with invasive breast cancer, and about 63,410 women were diagnosed with in situ breast cancer in the United States, according to estimates.

The following are some of the recognized breast cancer risk factors. Most incidences of breast cancer, on the other hand,

✉ T. R. Mahesh
trmahesh.1978@gmail.com

V. Vinoth Kumar
drvinothkumar03@gmail.com

V. Vivek
vivekforvivek@gmail.com

K. M. Karthick Raghunath
raguaut@gmail.com

G. Sindhu Madhuri
madhuri.ju2017@gmail.com

[1] Department of Computer Science and Engineering, Faculty of Engineering and Technology, JAIN University, Bangalore, India

[2] Department of Computer Science and Engineering, MVJ College of Engineering, Bangalore, India

cannot be attributed to a specific cause. As one becomes older, the chance of developing breast cancer also increases. Breast cancer is prevalent in about 80% of those above 50 years of age. A person who had breast cancer really in one Breast is more likely to have cancer in the other Breast. Having a close relative with a person having breast cancer, especially at a very young age, is associated with increased breast cancer risk (before 40). Having other relatives who have been diagnosed with breast cancer can further increase your risk. Specific genetic mutations, such as those in BRCA1 as well as BRCA2 genes, have been linked to an increased chance of getting breast cancer in the future. According to research, the risk of breast cancer seems to be related to reproductive and menstrual history. Early beginning menstruation is a higher risk factor (before age 12), Menopause with a late-onset (after age 55). Never having children, having children later in life, or not breastfeeding are all options. Hormones that are risk factors for breast cancer may be found in menopausal hormone therapy and specific birth control methods.

Some of the symptoms of breast cancer are a lump or a clump, changes in breast or nipple size or shape, Breast or nipple color changes, nipple inverted, The discharge of nipples, Breast swell or thickening, Consistent discomfort; Dimpling is a type of skin dimpling that occurs when the skin is, Irritated or flaky skin. Mammography or a portable cancer diagnostic instrument can be used to detect it early during a screening test. The breast cancer tissues change as the disease progresses, which can be connected to the cancer stage. The breast cancer stage (I–IV) indicates how far cancer has extended in that patient. Different stages are discovered using statistical indications like tumour size, distant metastases, lymph node metastasis. Patients must have breast cancer surgery, chemotherapy, radiation, endocrine therapy to stop cancer from further spreading. Breast cancer occurs in several forms, the most frequent ductal carcinoma in situ (DCIS) and invasive carcinoma. The other conditions, such as phyllodes, tumours and angiosarcoma, are quite rare. There are a variety of algorithms for categorizing breast cancer outcomes. Fatigue, headaches, pain and numbness (peripheral neuropathy), bone loss, and osteoporosis are all side symptoms of breast cancer. There are numerous methods for classification as well as prediction of this disease. Breast cancers may be classified using a unique ensemble classification algorithm proposed in this study. We used SVM, LR, RF, DT, NB and KNN as base learners for the proposed Blended Ensemble classification model. In addition, Wisconsin Breast Cancer Dataset from Kaggle and the UCI machine learning repository are used to assess the performance of the suggested technique. The research's purpose is to detect and categorize malignant and benign patients and improve prediction accuracy.

2 Related work

Machine learning and its associative techniques in healthcare have been recognized as critical in improving patient outcomes and wellbeing. Using Logistic Regression (Telsang and Hegde 2020) an accuracy of 96.4% has been achieved. In this study (Akbugday 2019) SVM along with KNN has been used to classify the Breast Cancer and achieved 96.85% accuracy. RF (Keles 2019) was employed and achieved 92.2% accuracy. To determine the optimal classifier in the dataset of breast cancer (Delen et al. 2005) analysis was done with respect to the performance of the classifiers of Nave Bayes, SVM-RBF kernel, DT, RBF neural networks, and basic CART. ADABOOST was employed and it outperformed Random Forest by 97.5 percent. In this work (Assiri et al. 2020) ensemble methods were employed to obtain 96.25% accuracy than earlier investigations (Asri et al. 2016) where back propagation approach was used with 96.2% accuracy. The results revealed that the SVM-RBF kernel gives better performance compared to other classifiers, scoring 96.84% accuracy in Wisconsin dataset breast cancer. They employed SVM, KNN, Random Forest, Nave Bayes, and ANN as classification algorithms.

The genetic algorithm surpasses weighted average approaches in a comparison of particle-swarm optimization, deferential evolution (DE), as well as genetic algorithm (Chauhan and Swami 2018). Another comparison is made between the traditional ensemble approach and the GA-based weighted-average technique, and the genetic algorithm based weighted-average technique is found to outperform (Agarap 2018). Random Forest and Boosted Trees, two categorization models, had similar accuracy (Gupta and Shalini 2018). As a result, the most accurate classifier is being used to identify the tumor early on, allowing the cure to be discovered. (Aslan et al. 2018). Classification, regression, and clustering are examples of data mining techniques (Olson and Delen 2008) that assist us in obtaining useful information about patients having breast cancer. These algorithms (Li et al. 2017) include a training dataset that may be used to determine the likelihood of predicting various types of breast cancer (Sarveshvar et al. 2021).

Data mining (DM) is the process of extracting useful information out of a huge dataset. DM techniques and functions such as ML, statistics, database, fuzzy set, data warehouse, and neural network aid in the diagnosis and prognosis of various cancer diseases (Gupta and Chandra 2020) such as prostate cancer, lung cancer (Delen 2009), and leukemia (Shahbaz et al. 2012). Traditional cancer detection methods are based on "the gold standard" procedure, which includes 3 tests: clinical evaluation, pathology testing and radiological imaging (Shashikala et al. 2021). This traditional method, which is based on regression, detects the existence of cancer, whereas new ML methods as well as algorithms are completely based on creation of model. In its training and testing stages, the model is meant to forecast unknown data and offers a satisfactory predicted outcome (Salod and Singh 2019). Preprocessing, feature selection or extraction, and classification are the three major methodologies used in ML (Eltalhi and Kutrani 2019). The key aspect of the ML method is feature extraction, which

aids in the diagnosis and also, prognosis of cancer. A process of this kind may differentiate between benign and malignant tumors (Shrestha et al. 2021). In the proposed system, we have implemented Ensemble learning using six supervised ML algorithms as base learners on the collected breast cancer dataset, and a substantial increase in the accuracy and recall is observed. Machine Learning models operate on various ideas by combining many models in improving the performance of the traditional model. The objective to introduce the ensemble learning and understanding basic algorithms by creating multiple models. Ensemble method gives accurate solutions than existing models. Ensemble method is used in our methodology which employs in predicting good accuracy results which fixes issues or any limitations as per research study. As per the research experiences ensemble methods is highly as a blended model.

3 Methodology

The likelihood of survival and the likelihood of cancer recurrence are highly dependent on medical treatment and the accuracy of the diagnosis. Arbitrarily extracted information was used in this investigation, with the ratio of 70:30 split between training and testing data. The model was trained utilizing training sets, and its effectiveness was tested via test data. The dataset consists of 10 features or attributes whose values will determine whether the person is likely to affect with breast cancer or not, and the dataset has 143 instances. The target or output variable is a binary variable that is either malignant or benign. The different phases present in the process is depicted in Fig. 1.

The first step is to gather the data which are required for pre-processing in order to improve the quality of data by using data cleaning techniques, data transformation and data normalization. Data pre-processing is a DM technique that entails transforming the raw data into a suitable format which can be understood. In reality, the real-world data is inadequate, not consistent, and deficient, and it is almost always riddled with inaccuracies. In the second phase, Data pre-processing, a technique of DM for filtering data into a useful format as real-world data is almost always available in a variety of formats. It isn't available in the way that is needed, so it needs to be filtered in a way that one can understand. For data preprocessing, the standardization method is employed to transform the dataset into a usable format. Feature selection, which is called as attribute selection in ML and statistics, is a methodology of choosing a subgroup of relevant attributes for use in the creation of model. The main step involved in this methodology is feature selection

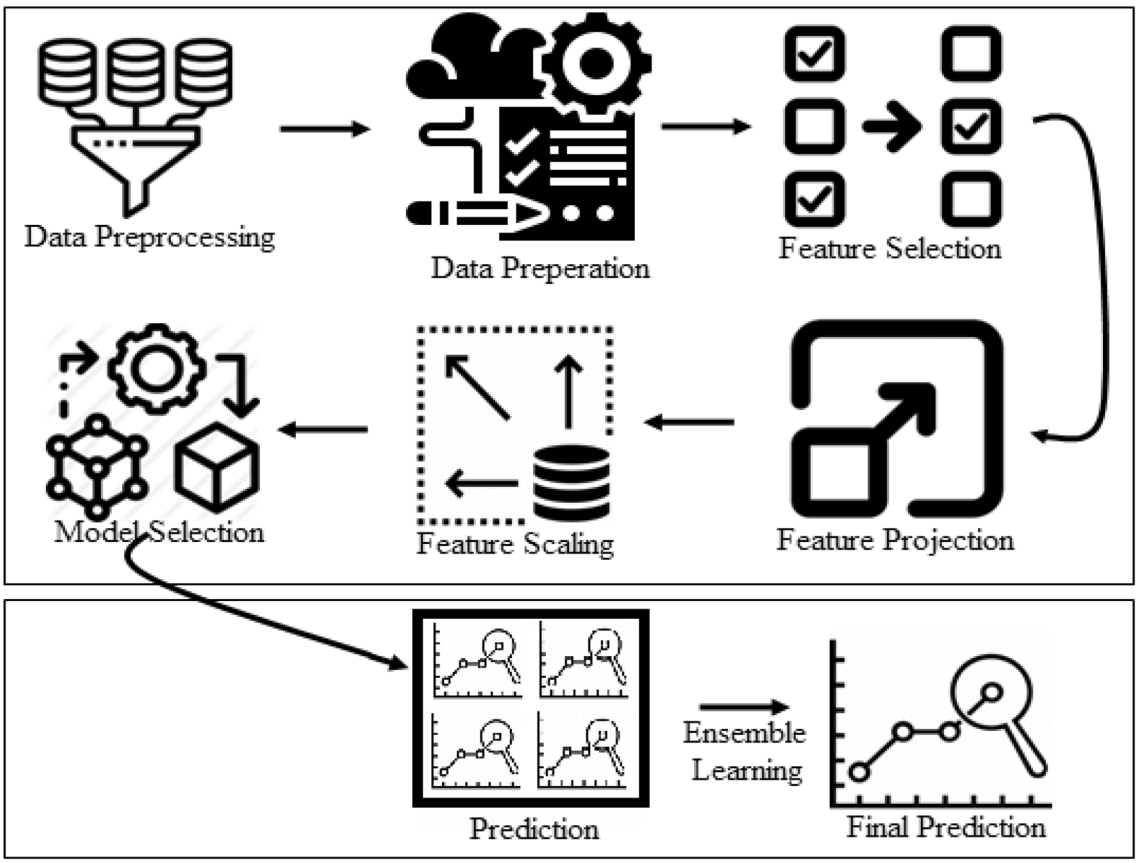


Fig. 1 Process of proposed methodology

and scaling method as it is mainly used to standardize maximum and minimum range of the independent variables or data consisting of important features.

Breast cancer dataset images consists of normal, malignant, benign is shown in Fig. 2. The dataset taken from the kaggle consists of 10 independent variables namely, Sample code number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses and one dependent or output variable. However, the first feature Sample code number is not considered for processing as it does not have any significance. There are about 143 samples in the dataset. Data is summarized using the python functions like dataset shape and dataset.head (5) (Shrestha et al. 2021c). Segregation of data is done where in X contains all the input variables and Y contains the class label that is the output variable. Next, Data splitting is done where in 70% of the entire dataset is taken for training purpose and 30% of the dataset is test data.

Depending on the nature of correlations between the features in the dataset, both linear and nonlinear reduction strategies can be applied (Kharya and Soni 2016). However, the majority of ML algorithms will compute the Euclidian distance between two data points. All characteristics must be brought to the same magnitude level. This can be accomplished through scaling (Huang et al. 2020). The model can learn from the training set and use it to predict the result of future data. They're divided into two categories: regression and classification (Khan et al. 2019). Machine learning is used to answer queries in the prediction phase. So, inference, or prediction, is the point when we get to answer some questions. The following six ML algorithms have been implemented to have comparison with respect to performance metrics.

3.1 Regression analysis

Regression analysis (RA) is a set of procedures of statistics for relationships estimation between a variable that is dependent and one or more variables that are independent. It is being used to determine how actually strong a relationship between variables is as well as to predict how they are going to interact in the future.

3.2 Bayesian statistics

It is a method of data analysis which is based upon Bayes' theorem (Dhiman et al. 2021). To estimate the posterior distribution, background knowledge is described as a prior distribution and paired with observational data in the form of a probability function. Bayesian statistics is a method for analyzing data and estimating parameters based on Bayes' theorem.

3.3 K-nearest neighbors (KNN)

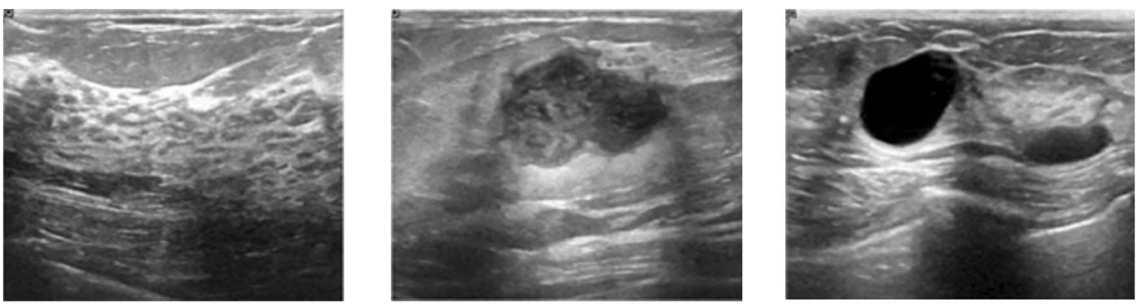
The supervised machine learning algorithm k-nearest neighbors (KNN) is a simple, easy-to-implement technique that may be used to address both classification and regression issues. The KNN algorithm believes that objects that are similar are close together. To put it another way, related items are close together. In our methodology, K-value is considered as 4.

3.4 Decision trees classifier

It is one of the most extensively used and practical approaches in Machine Learning since it are simple to use and interpret. Classification as well as regression both problems can being solved with decision trees (Kumar et al. 2021). The name suggests that it use a tree-like flowchart to display the predictions that result from a sequence of splits which are feature-based.

3.5 Random forest (RF) classifier

RF works by using the training data to create several decision trees. In the case of classification, every tree suggests output as a class, also the class with the maximum number of outputs is chosen as the final outcome (Shahbaz et al. 2012).



(a) Normal (b) Malignant (c) Benign

Fig. 2 Breast Cancer Sample images from kaggle dataset

We must specify the number of trees we wish to build in this algorithm. RF is such a technique for aggregating or even bagging bootstrap data. This method is being used to reduce an important parameter called variance in the outcomes. Ensemble approaches strive to improve model predictability by merging multiple models into a single, highly dependable model. Boosting, bagging, and stacking are the most prevalent ensemble approaches. Ensemble approaches are particularly well suited to regression and classification, where they reduce bias and variance while increasing model accuracy. Random forests, also known as random decision forests, are an ensemble learning method for classification, regression, and other problems that works by training a large number of decision trees. For classification tasks, the random forest's output is the class chosen by the majority of trees.

3.6 Support vector classifiers

Support vector machines (SVM) is supervised learning algorithms for classification, regression, and detection of outliers. In high-dimensional spaces, SVM works well. When the number of dimensions exceeds the number of samples, it is still effective. It is memory efficient because it uses a subset of training points in the decision function. Both dense and sparse sample vectors are accepted as inputs by scikit-support learning vector machines. To use an SVM to create predictions for sparse data, however, it must have been fitted on sparse data. In this methodology, the sigmoid kernel of SVM is used with the value of alpha as 0.3.

3.7 Ensemble learning technique

Unlike stacking, the predictions are made on the holdout set only. The holdout set and the predictions are used to build a model which is run on the test set. Steps are as follows: (1) The training set is further split into training and validation sets and models are fitted on the training set. (2) The predictions are made on the validation set and the test set. (3) The validation set and its predictions are utilized as features to fabricate another model used to make final predictions on the test and meta-features. In this proposed approach, the Ensemble classifier is blended by using Random Forest classifier. The EL model is depicted below:

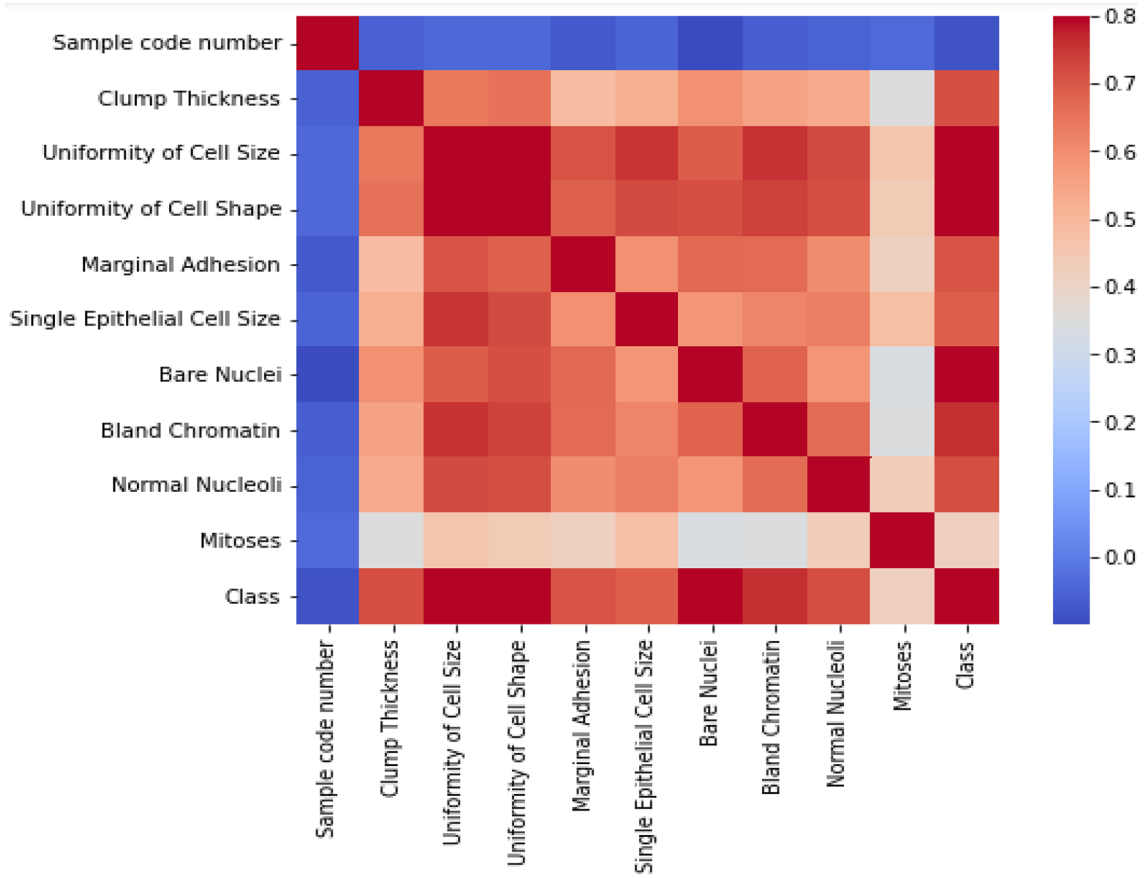


Fig. 3 Heatmap for the breast disease dataset

Algorithm 1: Ensemble Learning Model

Input: Training and Testing Datasets. $T_r = \{a_i, b_i\} \quad ; i = 1 \rightarrow k$, $T_D = \{a_j, b_j\} \; ; j = 1 \rightarrow k$
Output: Ensemble Prediction

Step_1: $T_r \rightarrow \{B_i\} \wedge \{h_i\}$ *where*, $B_i = \{a_i, b_i\}_{i=1 \rightarrow x}$ *and* $h_i = \{a_i, b_i\}_{i=1 \rightarrow y}$
Step_2: Learning Base Model, b_n
$\forall: n = 1$ *to* N *Do*
Learn: $\langle b_n | B_i \rangle$
End $\forall$
Step_3: Prediction on h_i
$\forall: n = 1$ *to* y *Do*
$h_i = [(a_i), \{h_1(a_1), \ldots, h_i(a_i)\}]$
End $\forall$
Step_4: Prediction on T_D
$\forall: n = 1$ *to* l *Do*
$T_D = [(a_i), \{h_1(a_1), \ldots, h_i(a_i)\}]$
End $\forall$
Step_5: Learning Meta Model, h_i
$\forall: n = 1$ *to* y *Do*
Learn: $\langle V_n | h_i \rangle$
End $\forall$
Step_5: Prediction on T_h *using* V_n
$T_D = V \times [(a_i), \{h_1(a_1), \ldots, h_i(a_i)\}]$

4 Comparison of results and discussions

The dataset used is Breast Disease UCI taken from the Kaggle and UCI machine learning repositories. It consists of 10 input variables and the heatmap is depicted in Fig. 3. Color-coded systems are used to create heat maps, which are graphical representations of data. Heat Maps are primarily used to better represent the amount of locations/events within a dataset and to guide users to the most important sections on data visualizations. A heatmap is a 2-dimensional graphical representation of data which uses colors to depict the individual values in a matrix.

The receiver operating characteristic curve (ROC) is a graph that depicts a classification model's performance across all classification levels. There are two parameters that are plotted on this curve: Firstly, true positive rate (TPR) is a measure of how often something is true, secondly False Positives is a measure of how often something is true. The TPR is calculated and plotted against the false positive rate for a single classifier at various thresholds to form the ROC curve which is depicted in Fig. 4.

With the help of a confusion matrix, the accuracy was determined. Confusion matrices are created using a model's predictions on a data set. Also, one can grasp the strengths and shortcomings of the model by looking at this confusion matrix, and the comparison is done with alternative models to see which one is the best for prediction of heart disease as shown in the confusion matrix Fig. 5.

"Yes" and "no" are the two potential predicted classes. If we were forecasting the presence of breast cancer, "yes" would indicate that they have the condition, while "no" would indicate that they do not.

The proposed model has been evaluated with respect to different performance measures namely precision, f1-measure and recall. The results are shown in the Fig. 5. The fraction of correctly categorized events among those classified as positive is measured by precision. Precision gives what percentage is truly positive out of the entire positive predicted. Precision is the ratio of true positives (TPs) to the total of TPs and true negatives (TNs). When the costs of False Positive (FNs) are large, precision is a good metric to use. The precision is computed using the Eq. 1.

$$Precision = \frac{TPs}{TPs + FPs} \tag{1}$$

The ratio of accurately predicted positive instances divided by the total number of positive examples predicted is used to compute it. The precision of LR, NB, KNN, RF, DT and SCM classifiers are 94.33%, 88.67%, 97.91%, 94.54%, 94.54% and 96.29% respectively.

Recall gives what percentage are predicted positive, out of the total positives. Recall is the ratio of TPs to the total of

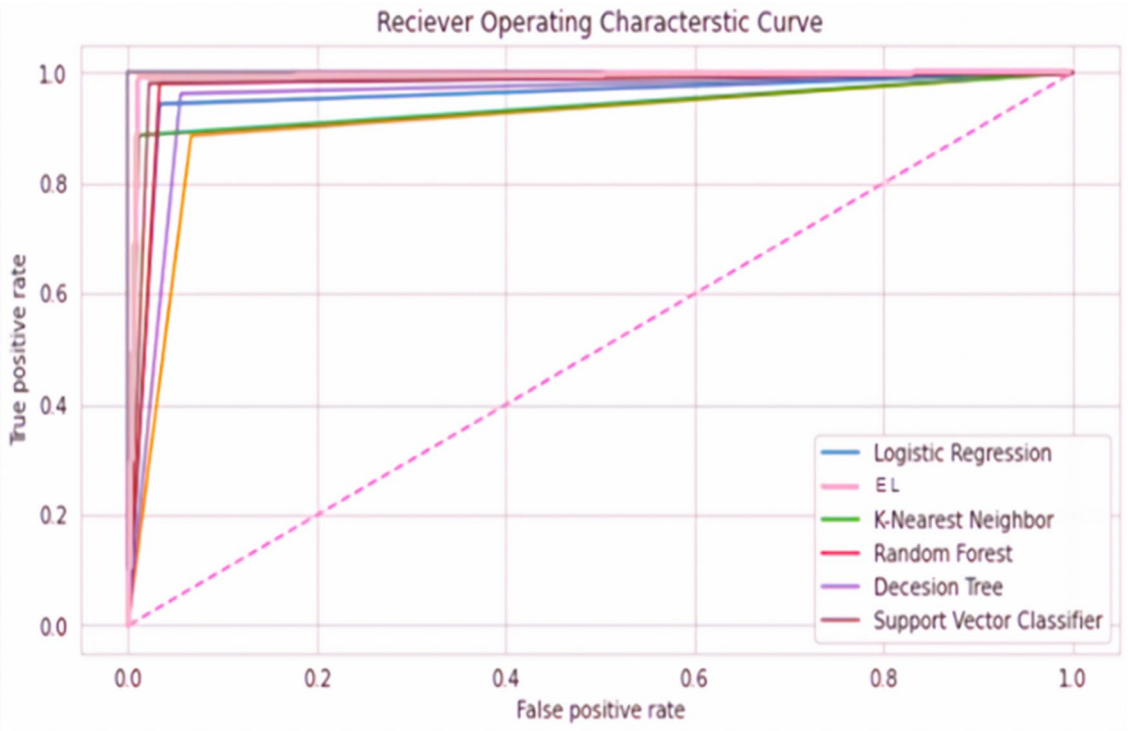


Fig. 4 ROC curve

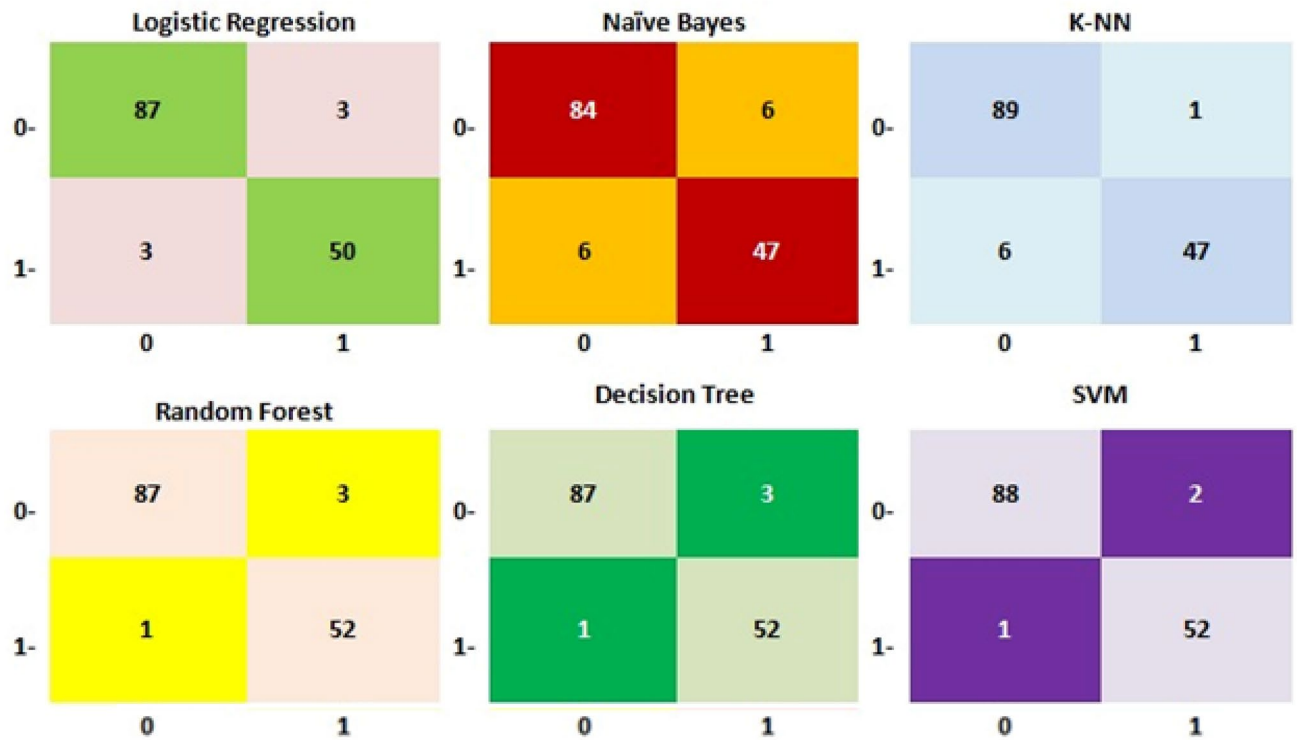


Fig. 5 Confusion matrix of six different ML Algorithms

TPs and FNs. Recall is same as true positive rate. Recall is a metric that measures how many correct positive predictions were produced out of all possible positive predictions. The recall is computed using the Eq. 2.

$$Recall = \frac{TPs}{TPs + FNs} \tag{2}$$

Unlike precision, which only considers the right positive predictions out of all positive predictions, recall considers the positive predictions that were missed. The recall of LR, NB, KNN, RF, DT and SCM classifiers are 94.33%, 88.67%, 88.67%, 98.11%, 98.11% and 98.11% respectively.

It is required to have both accuracy and recall to be one in a good classifier, which also means FP and FN should be zero. As a result, we require a statistic that considers both precision and recall. F1 score is the harmonic mean of precision as well as recall. Precision and Recall works very well when the dataset is balanced, however F1 score works very well when the dataset is imbalanced. When compared to accuracy performance indicators, the confusion matrix, precision, recall, and F1 score provide better insights into the prediction. The F1-score is a measure that takes precision and recall into account and is defined as follows as shown in Eq. 3.

$$F1\ Score = 2 * \frac{precision * recall}{precision\ +\ recall} \tag{3}$$

The f1-score of LR, NB, KNN, RF, DT and SCM classifiers are 94.33%, 88.67%, 93.06%, 96.29%, 96.29% and 97.19% respectively. The Fig. 5 shows the performance metrics of the different ML models. The performance of these six implemented ML algorithms with respect to recall, precision and f1-score are shown in Fig. 6.

The number of correctly classified data instances divided by the total number of data instances is known as accuracy. The Accuracy is computed using the Eq. 4. SVM and Random Forest are found to be performing very good compared to other algorithms. It is quite natural that RF gives its best as it is an ensemble method.

$$\text{Accuracy} = \frac{TNs + TPs}{TNs + TPs + FPs + FNs} \tag{4}$$

The accuracy of the different ML algorithms is also computed. The number of correctly classified data instances divided by the total number of data instances is known as accuracy.

The accuracy of LR, EL, KNN, RF, DT and SVM classifiers are 96.32%, 98.14%, 94.11%, 95.22%, 97.2% and 93.46% respectively. Out of six ML algorithms when evaluated with respect to several performance metrics, like precision, recall and F1 score, it is observed that EL, RF as well as LR has been evaluated with a very good accuracy compared to other classifiers. The comparison of the accuracy of all the six ML algorithms is depicted in Table 1.

Figure 7 illustrates the processing time required for several classification techniques. In comparison to other methods, EL techniques take longer to process data due to the blending process of base learners. CG (conjugate gradient) optimization was utilized for primary LR with compliance. L2 regularization and Alpha = 0.0001 were being used by SVM to get the best learning rate. Using Gini impurity, a DT was divided into smaller trees which consume optimal time for processing. It was enlarged until all leaves were pure or included fewer than two nodes, whichever came first.

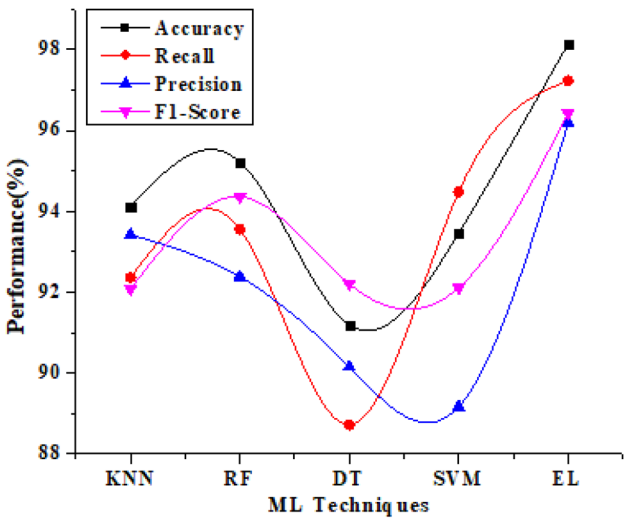


Fig. 6 Performance evaluation of different ML techniques in breast cancer classification

5 Conclusion

Breast cancer, if detected early, can save the lives of thousands of thousands women. These programs assist patients and clinicians in gathering as much information as possible in the real world. The objective is to determine the best

Table 1 Performance of various ML techniques in classification of breast cancer

ML techniques	Accuracy	Recall	Precision	F1-score
LR	96.32	89.14	95.17	94.27
KNN	94.11	92.36	93.42	92.08
RF	95.22	93.56	92.38	94.36
DT	91.19	88.72	90.15	92.21
SVM	93.46	94.48	89.17	92.11
EL	98.14	97.23	96.18	96.43

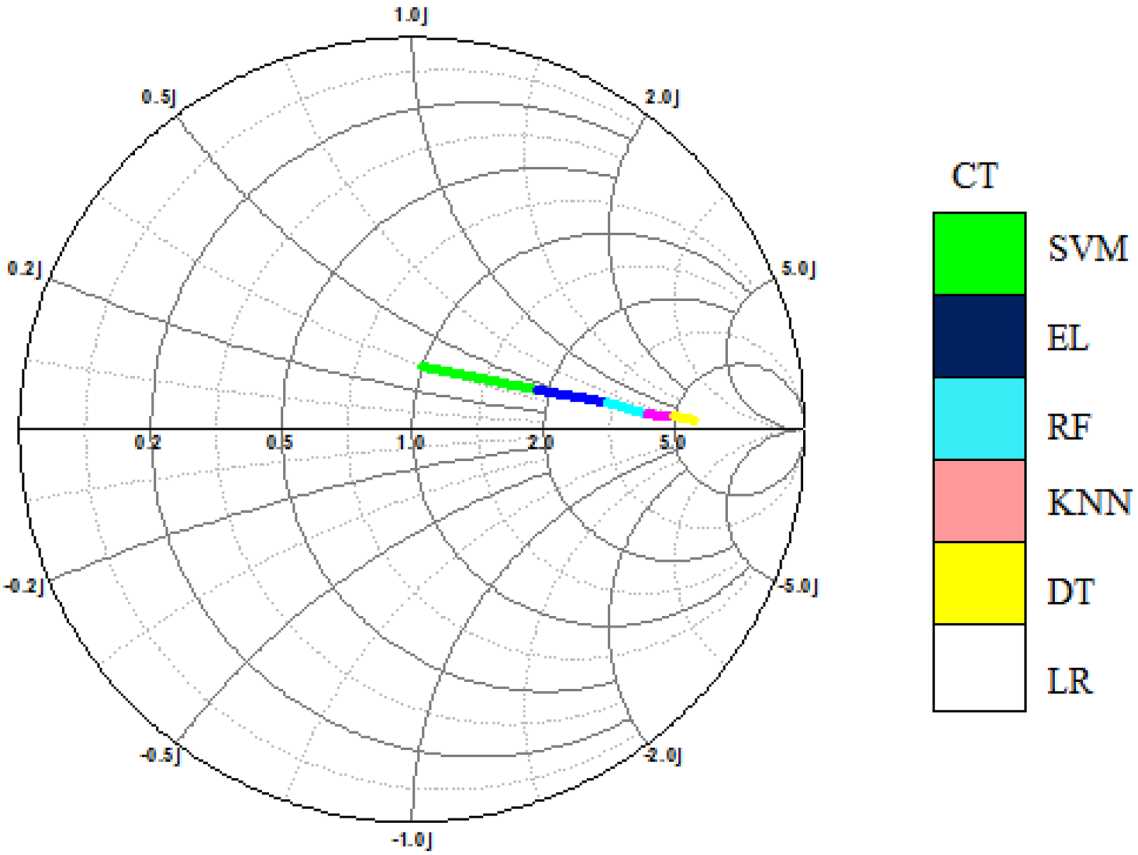


Fig. 7 Computation analyses of ML techniques

algorithm for forecasting the occurrence of breast cancer quite accurately. The main goal of this article is to highlight all of the previous and existing studies of ML algorithms that have been used to predict breast cancer. Also, this paper provides all of the necessary and required information to the research beginners those who want to examine the ML algorithms in order to gain a deep learning foundation. In summary, Blending EL exhibits their power from the perspective of effectiveness as well as efficiency with respect to precision, accuracy, f1-score and recall. More research in this area is needed to improve the classification techniques' performance so that they can forecast on more number of variables. The effort is made to figure out how to parameterize our categorization approaches so that high accuracy can be achieved. Investigation of a variety of datasets to see how Machine Learning techniques may be utilized to better characterize Breast Cancer is still happening. The aspiration is to achieve 100 percent accuracy by reducing error rates.

Funding No funds has been recieved to carry out this research work.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Human and animal rights No animals and humans participated in this research.

Informed consent No consent.

References

Agarap AFM (2018) [ACM Press the 2nd international conference—Phu Quoc Island, Viet Nam (2018.02.02–2018.02.04)] Proceedings of the 2nd international conference on machine learning and soft computing—ICMLSC '18—"On breast cancer detection", pp 5–9

Akbugday B (2019) 2019 Medical Technologies Congress (TIPTEKNO)—Izmir, Turkey (2019.10.3–2019.10.5)] 2019 Medical Technologies Congress (TIPTEKNO)—"Classification of Breast Cancer Data Using Machine Learning Algorithms", pp 1–4

Aslan MF, Celik Y, Sabanci K, Durdu A (2018) Breast cancer diagnosis by different machine learning methods using blood analysis data. Int J Intell Syst Appl Eng 6(4):289–293

Asri H, Mousannif H, Al Moatassime H, Noel T (2016) Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput Sci 83:1064–1069

Assiri AS, Nazir S, Velastin SA (2020) Breast tumor classification using an ensemble machine learning method. J Imaging 6(6):39

Breast Cancer (2018) Statistics, Approved by the Cancer.Net Editorial Board, 04/2017. [Online]. http://www.cancer.net/cancer-types/breast-cancer/statistics. Accessed 26 Aug 2018

Chauhan P, Swami A (2018) Breast cancer prediction using genetic algorithm based ensemble approach. In: 2018 9th international conference on computing, communication and networking technologies (ICCCNT), 2018, pp 1–8

Delen D (2009) *Analysis* of cancer data: a data mining approach. Expert Syst 26(1):100–112

Delen D, Walker G, Kadam A (2005) Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 34(2):113–127

Dhiman G, Vinoth Kumar V, Kaur A, Sharma A (2021) DON: deep learning and optimization-based framework for detection of novel coronavirus disease using x-ray images. Interdiscip Sci Comput Life Sci 13:260–272

Eltalhi S, Kutrani H (2019) Breast cancer diagnosis and prediction using machine learning and data mining techniques: a review. IOSR J Dent Med Sci 18(4):85–94

Gupta MK, Chandra P (2020) A comprehensive survey of data mining. Int J Inf Technol 12:1–15

Gupta P, Shalini L (2018) Analysis of machine learning techniques for breast cancer prediction. Int J Eng Comput Sci 7(05):23891–23895

Huang Q, Chen Y, Liu L, Tao D, Li X (2020) On combining bi-clustering mining and AdaBoost for breast tumor classification. IEEE Trans Knowl Data Eng 32(4):728–738

Keles MK (2019) Breast cancer prediction and detection using data mining classification algorithms: a comparative study. Tehn Vjesn Tech Gazette 26(1):149–155

Khan S, Islam N, Jan Z, Din IU, Rodrigues JJ (2019) A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognit Lett 125:1–6

Kharya S, Soni S (2016) Weighted naive bayes classifier: a predictive model for breast cancer detection. Int J Comput Appl 133(9):32–37

Kumar D, Swathi P, Jahangir A, Sah NK, Vinothkumar V (2021) Intelligent speech processing technique for suspicious voice call identification using adaptive machine learning approach. In: Advances in computational intelligence and robotics, pp 372–380

Li L, Wu Y, Ou Y, Li Q, Zhou Y, Chen D (2017) [IEEE 2017 IEEE 28th annual international symposium on personal, indoor, and mobile radio communications (PIMRC)—Montreal, QC, Canada (2017.10.8–2017.10.13)] 2017 IEEE 28th annual international symposium on personal, indoor, and mobile radio communications (PIMRC)—"Research on machine learning algorithms and feature extraction for time series", pp 1–5

Olson DL, Delen D (2008)."Advanced Data Mining Techniques", Springer, 2008, ISBN: 978–3–540–76917–0.

Park SH, Han K (2018) Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 286:171920

Salod Z, Singh Y (2019) Comparison of the performance of machine learning algorithms in breast cancer screening and detection: a protocol. J Public Health Res 8(3):jphr-2019

Sarveshvar MR, Gogoi A, Chaubey AK, Rohit S, Mahesh TR (2021a) Performance of different machine learning techniques for the prediction of heart diseases. In: 2021a International conference on forensics, analytics, big data, security (FABS), 2021, pp 1–4

Shahbaz M, Faruq S, Shaheen M, Masood SA (2012) Cancer diagnosis using data mining technology. Life Sci J 9(1):308–313

Shashikala HK, Mahesh TR, Vivek V, Sindhu MG, Saravanan C, Baig TZ (2021b) Early detection of spondylosis using point-based image processing techniques. In: 2021b International conference on recent trends on electronics, information, communication & technology (RTEICT), pp 655–659

Shrestha P, Singh A, Garg R, Sarraf I, Mahesh TR, Sindhu Madhuri G (2021c) Early stage detection of scoliosis using machine learning algorithms. In: 2021c International conference on forensics, analytics, big data, security (FABS), pp 1–4

Telsang VA, nd Hegde K (2020) Breast cancer prediction analysis using machine learning algorithms. In: 2020 International conference on communication, computing and industry 4.0 (C2I4)