Comparative Analysis of RFE in Naïve Bayes Classification in Breast Cancer Detection

Jayant Dhingra[1] Abhinav Sharma[2]
[1,2]B. Tech Student
[1,2]Guru Tegh Bahadur Institute of Technology, New Delhi, India

Abstract— Breast Cancer is a major reason for increasing mortality rate among women. There has been exponential growth in the number of women suffering from the disease. An early and precise detection is the necessary as to prevent breast cancer successfully. Data mining techniques have an astounding potential to develop a system which can be utilitarian in cancer detection. To classify benign and malignant tumor we have used Naïve Bayes classification technique and further improved its accuracy by applying Recursive Feature Elimination. This paper is a comparative study on the implementation of Naïve Bayes Classifier and further improving its accuracy by applying Recursive Feature elimination on the same classifier, The classifier was implemented on the Wisconsin Diagnosis Breast Cancer (WDBC) dataset. Our experiments have shown that when Naïve Bayes was used for predictive analysis, it had an accuracy of 94.20% and an increased accuracy of 95.95% after applying Recursive feature elimination method.

Keywords: Breast Cancer, Naïve Bayes, Recursive Feature Elimination, Wisconsin Diagonosis Breast Cancer.

I. INTRODUCTION

Brest Cancer is the second most menacing cancer after the lung cancer. In the year 2018 according to the statistics provided by World Cancer Research Fund it is estimated that over 2 million new cases were recorded out of which 626,679 deaths were approximated. Of all the cancers, breast cancer constitutes of 11.6% in new cancer cases and come up with 24.2% of cancers among women [1]. For the detection of breast cancer, various techniques are used. Mammography is the most promising technique and used by radiologist frequently. Mammogram images are usually of low contrast and noisy. In breast mammography, bright regions represent cancer. The oncologist can also diagnose a tumor with physical examination of the breast and also by checking for inflammation of any lymph node in the armpit. The more popular of the above two mentioned techniques is the Magnetic resonance imaging (MRI) of the breast. If an abnormal area is seen on your mammogram, your doctor may request additional tests. A breast ultrasound uses sound waves to create a picture of the tissues deep in the breasts. An ultrasound can help your doctor distinguish between a solid mass, such as a tumor, and a benign cyst. If a doctor suspects breast cancer, they may order both a mammogram and an ultrasound. If both of these tests can't tell your doctor if you have cancer, your doctor may do a test called a breast biopsy. [3]

The applications of Machine Learning Algorithms in Medicinal conclusions is expanding moderately. The Machine Learning approach are used to prepare the proposed framework depend on include significance investigation and characterization calculations. Consequently, Machine Learning algorithms can be used for the classification of benign and malignant tumor.

The prior diagnosis of Breast Cancer can enhance the prediction and survival rate notably, so that patients can be informed to take clinical treatment at the right time. Classification of benign tumors can help the patients avoid undertaking needless treatments. Thus, the research is to be carried for the proper diagnosis of Breast Cancer and categorization of patients into malignant and benign groups. Machine Learning, with its advancements in detection of critical features from the complex datasets is largely acknowledged as the method in the prediction of breast cancer [1].

ML algorithms are to analyze any data set to extract data-driven model, prediction rule, or decision rule from the data set. Generally, in order to ensure the ML behave intelligently without human intervention, the system learns or extracts knowledge such as rules or patterns from a collection of input data or past experience.

A lot of ongoing research is being done on this area and by the help of various machine learning and data mining techniques for various breast cancer datasets. Many researchs have shown that these Machine Learning techniques have given good precision with a lot of fidelity in the prediction of the type of tumor along with the other various details about the tumor.

II. RELATED WORKS

Classification is a process of a predicting class in machine learning which is used to classify data point into a set of predefined classes or groups. Kathija et. al. proposed two classification approach on breast cancer dataset which is fetched from Wisconsin Diagnosis Breast Cancer (WDBC) dataset. For classification, they apply Support Vector Machine (SVM) and Naïve Bayes and compared both the techniques and got Naïve Bayes more accurate i.e 95.65% accuracy rate [2].

Ch. Shravya et. al. proposed three classification approach and compared their accuracy, they used Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbor (KNN) and applied these techniques on the dataset fetched from UCI repository and got SVM most accurate by receiving the accuracy of 92.7% [1].

Alireza Osareh et. al. applied SVM technique on two different datasets and received the accuracy of 98.80% and 96.33% [5].

In the paper by Sonali Nandish Manoli et. al. [6] has proposed two classification techniques i.e. Naive Bayes and Support Vector Machine (SVM) and further applied Principal Component Analysis (PCA) to reduce dimensions and increase its accuracy.

Vikas Chaurasia et. al. compared Naive Bayes, J48 and RBF network algorithms on a dataset to find the most

powerful predictive model and hence found out that Naive Bayes is the most accurate with the accuracy of 97.36% [7].

Naresh Khuriwal et. al. has proposed ANN and Logistic algorithm applied on Wisconsin Diagnosis Breast Cancer dataset and it shown that the ANN approach with logistic algorithm achieved 98.50% accuracy [8]. Moh'd Rasoul Al-hadidi et. al. has proposed two types of supervised learning models, which are Back Propagation Neural Network (BPNN) model and Logistic Regression (LR) model and compared their accuracy [9].

In another study Mohammad Sajjadieh et. al. proposed electromagnetic model which was based on Finite Difference Time Domain (FDTD). This model basically estimates the location of breast cancer tumours and this model has a higher accuracy than other signal processing estimation algorithms [10]. Another study by B.M. Gayathri et. al. proposed Relevance Vector Machine (RVM) method and compared it with other machine learning techniques and explained how RVM is better than other machine learning techniques [11].

Mamta Jadhav et. al. compared three machine learning techniques, which is, decision tree, logistic regression and random forest and out of which Random Forest classification found out to be most accurate with the accuracy of 98.6% [12]. Dursun Delen et. al. has compared three data mining methods that is, decision Tree, artificial neural networks and logistic regression and found out that decision tree is the most accurate with the accuracy of 93.6% [13].

Data Mining is the computer assisted process of digging through and analysing enormous sets of data and then extracting the meaning of the data. Data mining tools predicts behaviour and future trends. Classification is one of the tool and it widely helps in many field like banking, medical field and so on. Logistic Regression is one of the algorithm of classification and is very useful algorithm in creating predictive models and in classification method. Aung proposed a study on breast cancer and used logistic regression and found it very useful method in creating a predictive model [14].

III. METHODOLOGY

Naive Bayes algorithm in Machine Learning is based on the Bayes theorem which is named after a famous mathematician Thomas Bayes. Bayes theorem is a decision making tool and very convenient to use in medical field.

Bayes theorem is employed in clinical epidemiology to determine the probability of a particular disease in a group of people with specific characteristics.

The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

A Naïve Bayesian model is uncomplicated to make, without any iterative parameter estimation. Therefore, making it easier to use for massive information sets. Although the Naïve bayes classifier is comparitively simpler it does shockingly well and is widely used as it quintessentially outperforms a lot of sophisticated classifiers.

IV. ALGORITHMS

A. Naïve Bayes Algorithm

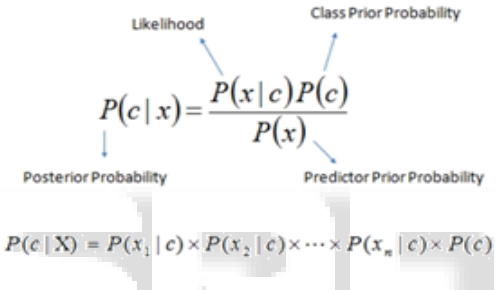
Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

P(c | x) is the posterior probability of class (target) given predictor (attribute).

P (c) is the prior probability of class.

P(x | c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor [15]



B. Recursive Feature Elimination

Finding optimal feature to use for Machine learning model technique can sometimes be a difficult task to accomplish. RFE is also a type of backward selection method and works on feature ranking system. First model is fitted on linear regression based on all variables, then it calculates variable coefficients and their importance after that it ranks the variable on the basis on linear regression fit and then remove low ranking variable in each iteration. Scikit package can do this automatically if you define the number of features you want to reduce it to.

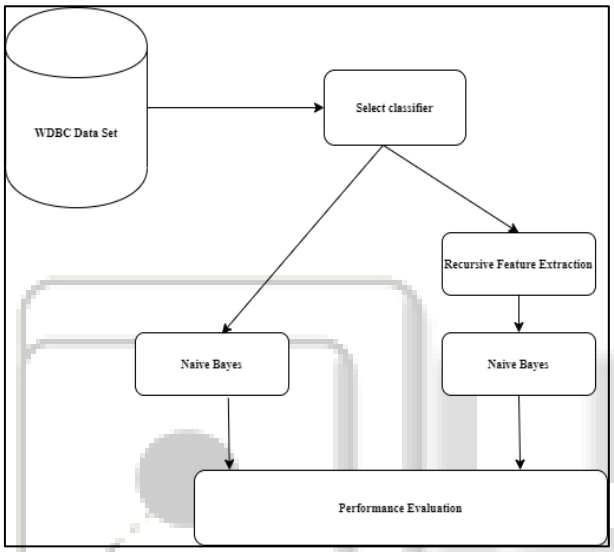
Recursive Feature Elimination (RFE) is a method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. In Machine Learning, selection of features and removing of all the unnecessary features plays a vital role. This helps us to increase the accuracy of any algorithm.

C. Breast Cancer Wisconsin dataset (Diagnostic)

The data used in this study are provided by the UC Irvine Machine Learning repository located in Breast Cancer Wisconsin sub – directory, filenames root: breast-cancer-Wisconsin having 569 instances, 2 classes (malignant and benign), and 32 real valued attributes. Class distribution: Benign: 359 (63.09%), Malignant: 210 (36.9%). Here, we have taken benign as 63.09% and malignant as 36.9% because it is better to take prevention than to cure, and therefore, large instances of benign patients have been taken for the study.

V. OUR IMPLEMENTATION

In this paper we proposed Naive Bayes and Recursive Feature Elimination (RFE) for prognosis and detection of breast cancer. We manually pre-processed the breast cancer data which we fetched from Wisconsin Breast Cancer Dataset (WBCD) and got 31 features out of 33 features and implement Naive Bayes on the selected features and got an accuracy of 94.200%. In order to further increase and compare the accuracy, we use Recursive Feature Elimination method for the pre-processing of the data and select 17 best features and then apply Naive Bayes to get the final accuracy and we have received the accuracy of 95.95%. Wisconsin Breast Cancer Dataset contains 569 rows and 33 columns. Achieved accuracy from Naive Bayes algorithm is further more increased by Recursive Feature Elimination method.



VI. PERFORMANCE EVALUATION

In this section we discuss the experiments and analysis set up, calculate performance and the effectiveness of our pro-posed solution approach. The following table and graph shows the increased accuracy after we applied RFE on the same dataset.

S.NO	Algorithms	No. of Columns	Accuracy (%)
1.	Naïve Bayes	31	94.200
2.	Naïve Bayes After RFE	17	95.95

Table 1: Comparison Table

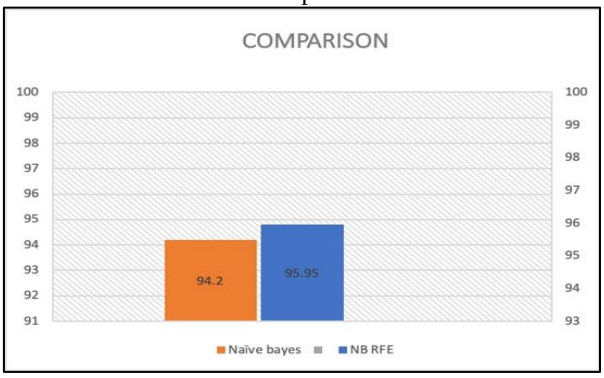


Fig. 1: Comparison Graph

VII. EVALUATION METRIC

To evaluate the effectiveness of the proposed solution ap-proach, standard confusion matrix is used. The following table shows the confusion matrix with each column of the matrix representing instances of predicted class while each row of the matrix rep-resenting instances of actual class. We execute Naïve Bayes classifier on test dataset of 569 patients and it classifies 359(353+6) as Benign and 210(193+17) as Malignant tumor. We evaluate the performance of our proposed solution ap-proach in terms of precision, recall, accuracy and f-score.

		Predicted	
		Is Benign	Is malignant
Actual	Is Benign	353(a)	6(b)
	Is malignant	17(c)	193(d)

Table 2: Confusion Matrix

1) Precision is the proportion of predicted relevant vide-os that were correct, calculated using the equation:

Precision = a / (a+c)

2) Recall is the proportion of relevant videos that were correctly identified, calculated using the equation:

Recall = a / (a+b)

3) Accuracy is the proportion of the total number of predictions that were correct, calculated using the equation:

Accuracy = (a+d) / (a+b+c+d)

4) F-Score is the weighted harmonic mean between pre-cision and Recall, calculated using the equation:

F-Score = Precision+Recall / 2 [16]

Precision	Recall	F-Score	Accuracy
0.9540	0.9832	0.9686	0.9595

Table 3: Performance Result

TABLE III shows the performance results of our proposed solution approach. Table III reveals that overall accuracy for cyberbullying detection is 95.95%

VIII. CONCLUSION

We present an approach based on Naïve bayes classification algorithms to classify benign and malignant tumor. Our experiments reveal that the proposed solution approach correctly able to identify the type of tumor with more than 95% accuracy. The proposed system has given the result with more accuracy compared to the existing system. We plan to investigate alternate and more efficient solutions to the breast cancer detection techniques with the help of other classification algorithms.

REFERENCES

[1] Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques", nternational Journal of Innovative Technology and Exploring Engineering, Vol.8, Issue-6, pp. 1106-1110, April 2019.

[2] Kathija and Shajun Nisha, "Breast Cancer Data Classification Using SVM and Naive Bayes Techniques", International Journal of Innovative Research in Computer and Communication Engineering, vol.4, Issue 12, pp. 21167-21175, December 2016.

[3] https://www.healthline.com/health/breast-cancer (accessed 14 February 2020).
[4] https://www.intechopen.com/books/breast-cancer-and-surgery/machine-learning-methods-for-breast-cancer-diagnostic (accessed 15 February 2020).
[5] Alireza Osareh and Bita Shadgar, "A Computer Aided Diagnosis System for Breast Cancer", International Journal of Computer Sciences, Vol.8, Issue-2, pp.233-240, March 2011.
[6] Sonali Nandish Manoli and Padma S.K, "Study and Analysis of Breast Cancer Data", International Journal of Engineering Research & Technology, Vol.5, Issue-12, 2017.
[7] Vikas Chaurasia, Saurabh Pal and BB Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques", Journal of Algorithms & Computational Technology, Vol.12(2), pp.119-126, 2018.
[8] Naresh Khuriwal and Nidhi Mishra, "Breast Cancer Diagnosis Using Adaptive Voting Ensemble Machine Learning Algorithm", Institute of Electrical and Electronics Engineers (IEEE), 978-1-5386-1138-8/18, 2018.
[9] Moh'd Rasoul Al-hadidi, Abdulsalam Alarabeyyat and Mohannad Alhanahnah, "Breast Cancer Detection using K-nearest Neighbor Machine Learning Algorithm", 9th International Conference on Developments in eSystems Engineering (DeSE, 2016.
[10] Mohammad Sajjadieh, Foroohar Foroozan, and Amir Asif, "Breast Cancer Detection using Time Reversal Signal Processing", IEEE 13th International Multitopic Conference, 2009.
[11] B.M.Gayathrin and Dr.C.P.Sumathi, "Comparative study of Relevance Vector Machine with various machine learning techniques used for detecting breast cancer", IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2016.
[12] Mamta Jadhav, Zeel Thakkar, Prof. Pramila M. Chawan, " Breast Cancer Prediction using Supervised Machine Learning Algorithms", International Research Journal of Engineering and Technology (IRJET), Vol.06, Issue-10, pp.851-854, Oct 2019.
[13] Dursun Delen, Glenn Walker and Amit Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods", ELSEVIER, pp.113-127, July 2004.
[14] Aung Pyae, "Classification of Breast Cancer using Supervised Machine Learning Algorithm", ResearchGate, May 2019.
[15] https://www.saedsayad.com/naive_bayesian.htm (accessed 15 February 2020).
[16] Mr. Shivraj Sunil Marathe, Prof. Kavita P. Shirsat, "Contextual Features Based Naïve Bayes Classifier for Cyberbullying Detection on YouTube", International Journal of Scientific & Engineering Research, Volume 6, Issue 11, November-2015.

