Naive Bayesian Machine Learning to Diagnose Breast Cancer

Arash Hooshmand (hooshmand@kth.se)
Kungliga Tekniska H□ogskolan

Research article

Keywords: Breast Cant Machine Learning, MI., DiA.105i8 Artificial htelligence AI, RNA Sequencing Transcriplomi. Otnics, Naiw, Baycsian, Classifitation, Clal..ifier Wbo Genonic Sequencing, WES

Posted Date: September 3rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-60997/v1

Noname manuscript No.
(will be inserted by the editor)

Naive Bayesian Machine Learning to Diagnose Breast Cancer

Effective Early Identification of Cancer by Machine Learning through Genome-wide RNA-Seq. Analyses

Arash Hooshmand

Received: date / Accepted: date

Abstract A novel MLAC (Machine Learning Against Cancer) method to distinguish between cancerous and noncancerous RNA genomic data is developed and tested with 100% accuracy on all healthy and cancerous Breast tissue samples. A Naive Bayesian ML (Machine Learning) system is trained using WES (Whole Exome Sequencing) data in a high-level i.e. normalized quantification of RNAs obtained from 1091 breast cancer samples' WES files from the TCGA (The Cancer Genome Atlas) and 179 healthy samples' WES data from the GTEx (Genotype-Tissue Expression) project. We could show that both sensitivity and specificity of the method in classification of cancerous and noncancerous cells is perfectly 100%.

Keywords Breast Cancer · Machine Learning · ML · Diagnosis · Artificial Intelligence · AI · RNA Sequencing · Transcriptomics · Omics · Naive Bayesian · Classification · Classifier · Whole Genome Sequencing · WES

1 Introduction

Breast cancer is one of the deadliest cancers in women. According to the American Cancer Society's 2019 biennial update, there were estimated 268600 new invasive cases of breast cancer and 41760 deaths because of it, and the importance of early diagnosis has repeatedly been emphasized. [1] Biologists have discovered many genes that are involved in specific cancers such as BRCA1 in breast cancer. Yet the signaling pathways and main reasons that cause most of cancers are unknown. Up until now, to avoid breast cancer, we know that alcohol and obesity are associated with an increased risk of breast cancer while

A. Hooshmand
Kungliga Tekniska Högskolan, Brinellvägen 8, 114 28 Stockholm, Sweden
Tel.: +46-73-9784199
Fax: +46-8-7900930
E-mail: hooshmand@kth.se

childbearing and breastfeeding have a protective effect. [2] [3] In diagnosis and cancer identification, histological examination is used that is a slow process and needs technical experts and suffers from large amount of variations among observers. In recent years, thanks to high throughput Omics technologies, we are no longer missing data but need novel methods and techniques to handle and analyze them; thus bioinformatics and computers have found a solid ground to contribute in Life Sciences. One of the most applicable approaches to benefit from Computer Science in Physiology and Medicine is utilization of AI (Artificial Intelligence) and ML to extract knowledge by computers out of Big Data generated by Omics technologies. [4] In this work, we have developed and trained a new ML-based system using general new generation of RNA Seq. data that can detect breast cancer even in very early stages, and hence will decrease the risk of mortality by early treatment.

ML is rapidly opening its position in medical and pharmaceutical sciences. Different models of ML have been tested in last few decades and have returned great results in different fields of medicine including but not limited to cancer identification. [5] [6] Naïve Bayes, Support Vector Machines, Random Forest, Logistic Regression, K Nearest Neighbors and Neural Networks (NN) are examples of general supervised ML algorithms that have reportedly been successful in different medical and pharmaceutical projects. [6] In this work we came up with a novel approach of applying ML for cancer detection that is effective and robust. Using our method, cancerous tissue can be identified easily in any stages, thus providing an opportunity to be controlled in time. This approach also offers a new direction for disease diagnosis while providing a new method to predict traits based on genomic information.

2 Method

In this project, we have used Naive Bayes algorithm from Sci-Kit Learn on 1270 samples from The Cancer Genome Atlas (TCGA) research network and the Genotype-Tissue Expression (GTEx) project portal and directly fed the genome data to the machine to do heavy statistical calculations on our high dimensional data. In below, different parts of the method are clarified.

2.1 Bayes' theorem

Bayes' theorem was proposed by the English Thomas Bayes in 1763 when he was trying to prove the existence of God by means of statistical inference. [7] Bayesian statistics are used in estimates based on anticipated subjective knowledge. Therefore, the implementations of this theorem adapt with use and allow combining the fusion of data from two or more different sources and expressing them in terms of likelihood. Naive Bayesian Classifier is an implementation of Bayes' theorem, with some additional simplifying hypotheses, which allow applying an independence hypothesis, between the predictor variables, hence "Naive" is added to the name of these implementations because

a naive Bayesian classifier assumes that the features of a class / object are not related to each other i.e. the presence of a particular feature is not related to the presence or absence of another. In this way each feature independently contributes to the probability of a given class. In return, Bayes Classifiers can easily be trained, require little data to train, and can classify big data quickly. Nevertheless naive Bayes classifiers are amazingly simple, they have worked quite well in many real-world situations, including our cancerous/healthy tissue classification. It required a small amount of training data and could be fast and accurate as reflected in the Results section. On the flip side, although naive Bayes is known as a decent classifier, it is known to be a bad estimator in a sense that one cannot rely on its parameters for extraction of feature importance. [9]

2.2 Model function

More formally, as shown by equations 1-6, Bayesian classifiers are, indeed, probabilistic classifiers using Bayes rule i.e.

$$P(A \mid B) = P(A)P(B \mid A)/P(B) \tag{1}$$

For example, A can be the prior probability of cancer and B the posterior probability of cancer; given positive cancer test result is the product of the prior times the sensitivity i.e. the chance of a positive result given cancer. Indeed, a naive Bayesian classifier accomplishes statistical inference based on maximum likelihood estimation i.e. setting the parameters of the probability distribution in a way that maximises the goodness of fit of a statistical model to the training data via joint probability distributions of the training samples. In technical words, the likelihood function describes a hyper surface whose peak, if it exists, is an arrangement of model parameters values and coefficients that maximize the probability of drawing the obtained sample. [8] In its more general form, according to Sci-kit Learn website documentation, Bayes' theorem states the following relationship, given class variable y and dependent feature vector x_1 through x_n:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)} \tag{2}$$

Using the naive conditional independence assumption that

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y) \tag{3}$$

for all i, this relationship is simplified to

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)} \tag{4}$$

Since $P(x_1, x_2, ..., x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y) => \hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y) \quad (5)$$

and we can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i \mid y)$; the former is then the relative frequency of class in the training set. The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(x_i \mid y)$. [9] GaussianNB implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \quad (6)$$

where the parameters σ_y and μ_y are estimated using maximum likelihood.

2.3 Feature selection

In ML supervised classification methods as well as in K-Means unsupervised clustering algorithm, the input data (the points) is viewed as a p-dimensional vector (an array or ordered list of p numbers). Then the classifiers more or less based on similar criteria e.g. in the Bayesian classifiers, the classifier looks for a hyper surface that maximizes the likelihood of drawing the sample, or in SVMs, it looks for a hyperplane that optimally separates the points of one class from the other, which eventually could have been previously projected to a higher dimensional space. There is wrong perceptions in the ML community that have prevented potential achievements, and we could get great results by violating fake red lines; one of them, for instance, is about the number of features such as "it is obviously impractical to select all of the point mutations as dimensions for the model because mass dimensions will increase the computation cost." As a result, researchers usually try to reduce by themselves the assumed learning pressure on the machines brought about by highly redundant dimensions and select a subset of features i.e. genes to reduce the number of features and dimensions. [10–12] A strength point of our work is that we consider ML as powerful advanced statistics tool doing heavy statistical analyses, that people themselves cannot do. As a result, we gave all the data corresponding to the WES as feature inputs to the ML at once and it returned almost perfect results quickly and precisely. We thought of 19627 different genes not as too many features but as different pixels of a less than 141*141-pixel photo and it was a very light task for the machine to analyze such a low resolution image and it took only seconds to classify the cancerous and noncancerous cells 100% precisely.

2.4 Model optimization and settings

We have used Gaussian Naive Bayes classifier from scikit-learn 0.23.1 with its default settings i.e. priors equal to None and var_smoothing equal to 1e-9 where var_smoothing is the portion of the largest variance of all features that is added to variances for calculation stability. Nevertheless naive Bayes classifiers are amazingly simple, they have worked quite well in many real-world situations, including our cancer/non-cancer classifier. They require a small amount of training data and can be extremely fast compared to other ML classifiers. On the flip side, although naive Bayes is known as a decent classifier, it is known to be a bad estimator. It means that one cannot rely on its parameters for extraction of feature importance. [9]

2.5 Model evaluation

Model evaluation produces measures to approximate a classifier's reliability. To distinguish between cancerous and noncancerous cells, since it is a binary classification, we use accuracy, precision, specificity, sensitivity, f1 score, several averaging techniques and ROC curve to evaluate the model. We, indeed, use Sci-kit Learn Metrics Classification Report that returns precision, recall and f1 score for each of two classes. In binary classification, recall of the positive class is called "sensitivity"; and recall of the negative class is "specificity". In what follows, the terms and derivations from confusion matrix such as accuracy, specificity, sensitivity, f1 score are given to review and compare:
Condition positive (P): the number of real positive cases in the data
Condition negative (N): the number of real negative cases in the data

True positive (TP) or hit
True negative (TN) or correct rejection
False positive (FP), false alarm or type I error
False negative (FN), miss or type II error

Sensitivity, recall, hit rate, or true positive rate (TPR):

$$TPR = TP/P = TP/(TP + FN) = 1 - FNR \tag{7}$$

Specificity, selectivity or true negative rate (TNR):

$$TNR = TN/N = TN/(TN + FP) = 1 - FPR \tag{8}$$

Precision or positive predictive value (PPV) is the ratio of the correctly labeled samples by our program to all labeled ones in reality.

$$PPV = TP/(TP + FP) = 1 - FDR \tag{9}$$

Precision can be calculated only for the positive class i.e. class 1 that shows cancer or can be evaluated for each one of the two classes independently treating each class as it is the positive class at time, and the latter is done in Sci-kit

Learn Metrics Classification Report as shown in table 1.

Negative predictive value (NPV):

$$NPV = TN/(TN + FN) = 1 - FOR \tag{10}$$

Miss rate or false negative rate (FNR):

$$FNR = FN/P = FN/(FN + TP) = 1 - TPR \tag{11}$$

Fall-out or false positive rate (FPR):

$$FPR = FP/N = FP/(FP + TN) = 1 - TNR \tag{12}$$

False discovery rate (FDR):

$$FDR = FP/(FP + TP) = 1 - PPV \tag{13}$$

False omission rate (FOR):

$$FOR = FN/(FN + TN) = 1 - NPV \tag{14}$$

Accuracy (ACC):

$$ACC = (TP + TN)/(T + N) = (TP + TN)/(TP + TN + FP + FN) \tag{15}$$

The harmonic mean of precision and sensitivity or f1-score (F1):

$$F1 = 2.PPV.TPR/(PPV + TPR) = 2.TP/(2.TP + FP + FN) \tag{16}$$

Since we are using Sci-kit Learn Metrics Classification Report to show the results as shown in table 1, we also describe the meaning of micro avg, macro avg and weighted avg. used in the report:

Micro-average of precision (MIAP):

$$MIAP = (TP1 + TP2)/(TP1 + TP2 + FP1 + FP2) \tag{17}$$

Micro-average of recall (MIAR):

$$MIAR = (TP1 + TP2)/(TP1 + TP2 + FN1 + FN2) \tag{18}$$

Micro-average of f-Score (MIAF) would be the harmonic mean of the two numbers above.

$$MIAF = 2.MIAP.MIAR/(MIAP + MIAR) \tag{19}$$

Macro-average of precision (MAAP):

$$MAAP = (Precision1 + Precision2)/2 \tag{20}$$

Macro-average of recall (MAAR):

$$MAAR = (Recall1 + Recall2)/2 \tag{21}$$

Table 1 Classification report

Summary	Precision	Recall	F1-score	Support
Class 0	1.00	1.00	1.00	22
Class 1	1.00	1.00	1.00	105
Micro avg	1.00	1.00	1.00	127
Macro avg	1.00	1.00	1.00	127
Weighted avg	1.00	1.00	1.00	127

Macro-average of f-Score (MAAF) would be the harmonic mean of the two numbers above.

$$MAAF = 2.MAAP.MAAR/(MAAP + MAAR) \tag{22}$$

Macro-average method is suitable to know how the system performs overall across different sets of data but should not be considered in any specific decision making because it calculates metrics for each label and finds their unweighted mean i.e. it does not take label imbalance into account, while in our case, the labels are highly imbalanced i.e. 1091 vs. 179. On the other hand, micro-average is a useful tools and returns measures for decision-makings especially when datasets vary in size because it calculate metrics globally by counting the total true positives, false negatives and false positives. Finally, Weighted-average, according to Sci-kit Learn documentation on f1-score metrics, calculates metrics for each label, and finds their average weighted by support (the number of true instances for each label). This alters "macro" to account for label imbalance; consequently, it can result in an F-score that is not between precision and recall.

3 Results

Genomic variation files for healthy people (179 persons) and cancer patients (1091 samples) were obtained from the Gtex Project and the TCGA online database. The results were just amazing because the system can detect all cancerous and noncancerous samples correctly and as seen in the classification report shown in table 1, the performance of the classifier is perfect with accuracy and precision of 100% and sensitivity and specificity of 1. In this classification, not only the accuracy is 100% but also the Receiver Operating Characteristic's Area Under Curve (ROC AUC) from prediction scores also would be 1 as seen in figure 1.

4 Discussion and Conclusions

The classifier did its task perfectly with no error, at least on our available data. There are yet some aspects to reflect on. Although most of TCGA Breast Cancer (BRCA) comprise white women's samples, but it contains samples of

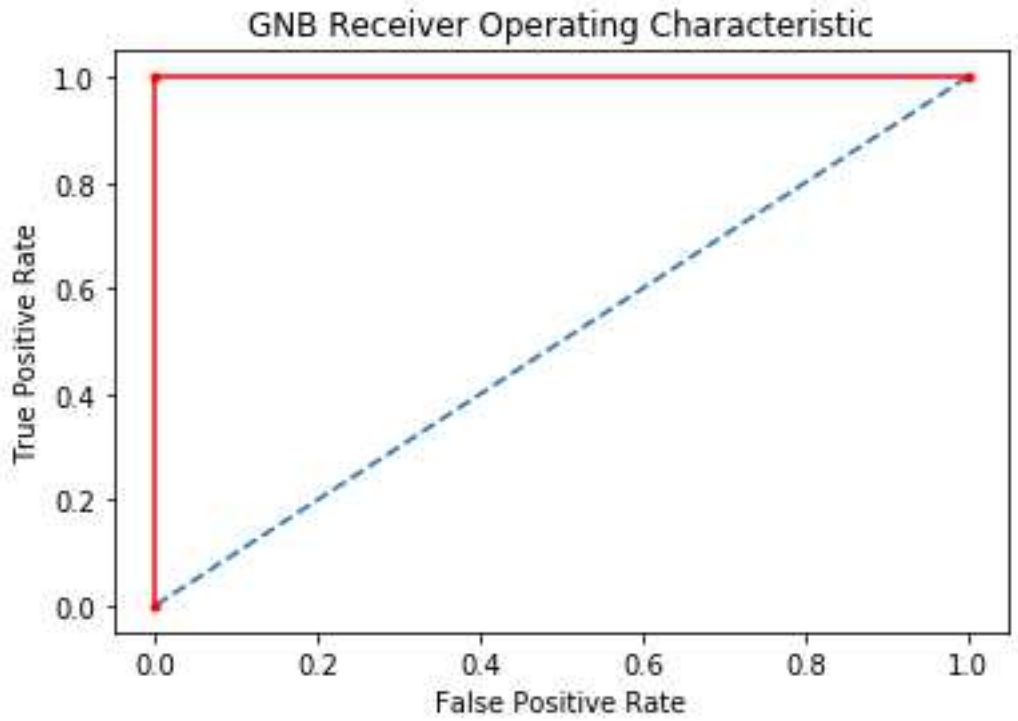

Fig. 1 ROC curve of Gaussian Naive Bayes classifier's performance on Breast Cancer data

Asian, black, and latin women as well as men. Since our method classifies all cancerous and non-cancer samples correctly using the information available in genomic variation, it means that the genetic signatures of cancer are detected universally without need to consider racial or sexual differences. The samples also from four different stages and the trained machine could detect cancerous samples from all stages correctly. This means even cancerous from earliest stages were distinguished correctly from noncancerous samples which is extremely important for effective treatments because cancer stage plays an important role in determining treatment options and patients' survival, as the earlier the diagnosis, the higher chance for the successful treatment.

Our work provided a new approach in application of ML on medical data that resulted in excellent classification between cancerous and noncancerous cells of the breast. In this work, we did not reduce the dimension of input data and left all the statistical analysis to the ML system and it could do its job very well and distinguished the cancerous samples from healthy cells almost perfectly. We even did not need to balance the number of samples of each class and it shows that the difference between two class is so much that providing hundreds of samples enables the machine to distinguish between two

categories perfectly without any mistake. We even did not need to pre-process the data obtained from Gtex and TCGA despite the fact that their data is not perfect and there are some rows of missing data for some genes quantities in some samples, yet the data provided by these two projects are fairly clean and reliable and it was enough for our classifier to be able to do its classification 100% correctly. This ML system is trained now to receive any new person's RNA-seq data and recognize if the patient's breasts are cancerous or not. It can detect the problem in different stages of cancer accurately; therefore, it can be helpful in early diagnosis of cancer. The limitation of our model is that it needs data of samples from organs and the involving labs should follow the same protocols on obtaining the transcriptomics data of 19627 genes as done by Gtex and TCGA on samples obtained from people's breasts. The New Generation RNA-seq protocols followed by Gtex and TCGA are well-known and standard. Thus the next work can be finding suitable biomarkers in the blood that can detect healthy people and patients only by their blood tests.

5 Declarations

5.1 Availability of data and materials

The data used in this project are publicly available on www.gtexportal.org and https://portal.gdc.cancer.gov/ and all ethical issues are strictly observed by them. This project does not need any extra personal/patient consent approval either because the data are normalized and does not reveal any private information and whatever necessary with respect to the law is observed by the institutes publishing them. All software can be available on github after paper approval.

5.2 Authors' contributions, competing interests and consent for publication

I, Arash Hooshmand, as the only author of this article have submitted it to the BMC Bioinformatics journal, hence I reiterate my consent to publish it in this journal. There is no competing interests and there is no need to any other consent approvals.

5.3 Funding and acknowledgement

The library of KTH Rotal Institute of Technology has Open Access publication agreements with BMC Bioinformatics and has accepted to pay for its publications. I also thank and acknowledge Houshmand family and their companies, especially Mr. Eng. GHolamAbbas Houshmand, Atash Houshmand, Shahab Houshmand, Shahin Houshmand and Shadab Houshmand for their financial support and contribution in the project.

References

1. DeSantis, Carol E., et al. Breast cancer statistics, 2019, CA: a cancer journal for clinicians 69.6: 438-451, (2019).
2. Momenimovahed, Zohre, and Hamid Salehiniya, Epidemiological characteristics of and risk factors for breast cancer in the world, Breast Cancer: Targets and Therapy 11: 151, (2019).
3. Lee, Priscilla Ming Yi, et al. Heterogeneous Associations Between Obesity and Reproductive-Related Factors and Specific Breast Cancer Subtypes Among Hong Kong Chinese Women, Hormones & Cancer (2020). Tsuji, S., and H. Aburatani., Machine Learning Applications in Cancer Genome Medicine, Gan to kagaku ryoho. Cancer & chemotherapy 46.3:: 423-426, (2019).
4. Nik-Zainal Abidin, S., Memari, Y., & Davies, H., Holistic cancer genome profiling for every patient, Swiss medical weekly, 150 w20158, 20158 https://doi.org/10.4414/smw.(2020).
5. Asri, Hiba, et al., Using machine learning algorithms for breast cancer risk prediction and diagnosis, Procedia Computer Science 83: 1064-1069, (2016).
6. Vamathevan, Jessica, et al. Applications of machine learning in drug discovery and development, Nature Reviews Drug Discovery, 18.6, 463-477, (2019). https://doi.org/10.1038/s41573-019-0024-5
7. Streiner, David L., Clinical medicine and the legacy of the reverend Bayes, International journal of clinical practice 73.4:e13323, (2019).
8. Myung, Jae., Tutorial on maximum likelihood estimation, journal of mathematical psychology, 47 (2003), 90100, (2002).
9. Zhang, Harry., The optimality of naive Bayes, AA 1.2, 3 (2004).
10. Pes, Barbara., Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains, Neural Computing and Applications: 1-23, (2019).
11. Sun, Yingshuai, et al., Identification of 12 cancer types through genome deep learning, Nature Scientific reports 9.1: 1-9 (2019).
12. Abeel, Thomas, et al., Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, Bioinformatics 26.3: 392-398, (2010).

Figures

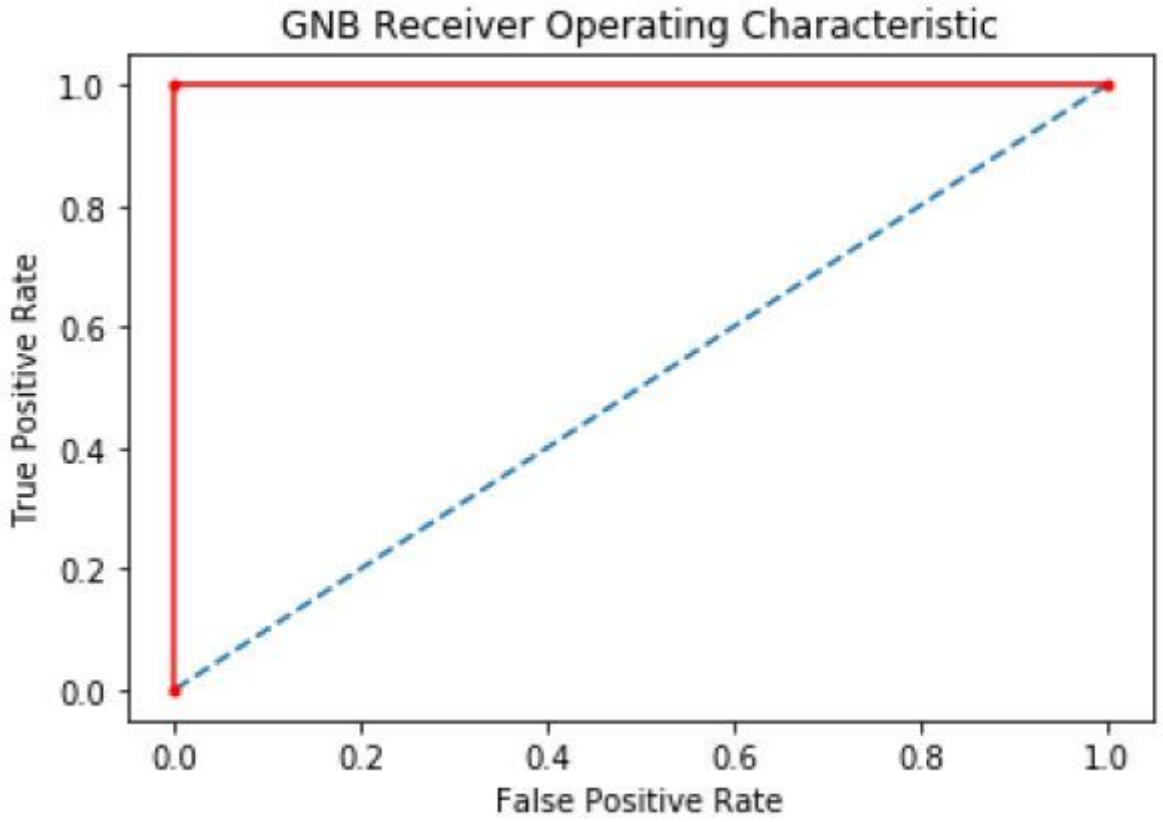

Figure 1

ROC curve of Gaussian Naive Bayes classier's performance on Breast Cancer data