Feature Selection for Breast Cancer Detection using Machine Learning Algorithms

Sreyam Dasgupta, Ronit Chaudhuri, Swarnalatha Purushotham,

Abstract: ***Cancer has been portrayed as a heterogeneous disease comprising of a wide range of subtypes. The early diagnosis of a cancer type is very important to determine the course of medical treatment required by the patient. The significance of classifying cancerous cells into benign or malignant has driven many research studies, in the biomedical and the bioinformatics field. In the past years researchers have been encouraged to use different machine learning (ML) techniques for cancer detection, as well as prediction of survivability and recurrence. What's more, ML instruments can be used to distinguish key highlights from complex datasets and uncover their significance. An assortment of these procedures, including Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Random Forest Methods (RVMs) and Decision Trees (DTs) has been usually used in cancer research for the development of predictive models, resulting in successful and exact decision making. Although it is obvious that the usage of machine learning techniques can enhance our comprehension of cancer detection, progression, recurrence and survivability, a proper level of accuracy is required for these strategies to be considered in the ordinary clinical practice. The predictive models talked about here depend on different administered ML strategies and on various input features and data samples. We have used Naïve-Bayes classifier, Neural Networks method, Decision Tree and Logistic Regression algorithm to detect the type of breast cancer (Benign or Malignant) and selection of features which are more relevant for prediction. We have made a comparative study to find out the best algorithm of the above four, for prediction of cancer type. With a high level of accuracy, any of these methods can be used to predict the type of breast cancer of any particular patient.***

Index Terms: ***Breast Cancer, Feature Selection, Logistic Regression, Naïve Bayes, Decision Tree, Neural Network, Machine Learning***

I. INTRODUCTION

It has been an attempt over last many years that how to detect and cure cancer. Cancer which can be of various types like breast cancer, lung cancer, throat cancer, blood cancer etc. is known to be the deadliest disease which still now have not got any cure. There are several levels of cancer from 1 to 6. However, on the good side if cancer is detected when it is in level 1 or 2 or at the very initial stage, there is a significant probability that it will get cured within a period of time. With the advent of new technologies in the field of medicine, we get new ideas of curing the disease with methods like machine learning. The problem for the cancer can be broadly classified into three types. Firstly, the problem is to predict whether a person in a particular stage of cancer has the chance to survive or not. Secondly the problem is to predict whether a person who has already encountered the disease in the past and got cured, has the probability of having the same disease in future. Thirdly the domain in which we are working includes the detection of cancer at the earliest stage. As they say prevention is always better than cure, we focus on all sorts of machine learning algorithms that can be used to detect the presence of cancer at an early stage. We all have the basic idea about what cancer is. Cancer is basically the malignant growth of tissues due to rapid cell division. Here we have a dataset specifying the various parameters of the patients and stating whether the tumour is malignant or benign. We are working in the domain of machine learning and in this paper we focus on the several types of standard algorithms like logistic regression, Naïve Bayes, Decision Tree, Neural Networks etc. to predict the presence of breast cancer for a given dataset. The data for over 500 patients are collected from websites like Kaggle for various parameters like Perimeter mean, Area Mean, Radius Mean etc. and we do a comparative study of all the algorithms to find out which algorithms gives the best accuracy results. We take the diagnosis of the dataset as the target variable and takes a certain percentage of the dataset to be the training set and the rest of it to be the testing set. This paper gives the results and comparisons for each type of learning algorithm in its plot. There are other methods related to profiling and circulating miRNAs that have been proven a promising class for cancer detection and identification. Various aspects regarding the prediction of cancer outcome based on gene expression signatures are discussed. These studies list the potential as well as the limitations of microarrays for the prediction of cancer outcome. Even though gene signatures could significantly improve our ability for prognosis in cancer patients, poor progress has been made for their application in the clinics. However, before gene expression profiling can be used in clinical practice, studies with larger data samples and more adequate validation are needed. In the present work only, studies that employed ML techniques for modeling cancer diagnosis and prognosis are presented.

II. LITERATURE SURVEY

Kourou, Themis Exarchos, Konstantinos Exarchos, Karamouzis and Fotiadis researched on machine learning applications in

Revised Manuscript Received on July 06, 2019.

Sreyam Dasgupta, Computer Science, Vellore Institute of Technology, Kolkata, India.
Ronit Chaudhuri, Computer Science, Vellore Institute of Technology, Kolkata, India.
Swarnalatha Purushotham, Computer Science, Vellore Institute of Technology, Vellore, India.

cancer prognosis and prediction in 2014. Their study proved that the integration of multidimensional heterogeneous data combined with the application of different techniques for feature selection and classification can provide promising tools for inference in the cancer prediction researches. Singhal and Tiwary proposed a new method for detection of skin cancer using Artificial Neural Network in 2015. First, they used wavelet transform for feature detection and then those features were used to train and test he neural network. This method successfully detects skin cancer from images with an accuracy of 92% with BPNN and 88% with RBFNN using a haar wavelet. Rana Bhat, Vishwanath and Li researched on detecting cancer through gene expressions using deep generative learning in 2016.They proposed a model for detection and classification of inflammatory breast and prostate cancer. The proposed model utilized cDNA microarray gene expressions to gauge its efficacy. Based on deep generative learning, the tuned discriminator and generator models, D and G respectively, learned to differentiate between the gene signatures without any intermediate manual feature handpicking, indicating that much bigger datasets can be experimented on the proposed model more seamlessly. Bashiri, Ghazisaeedi, Safdari, Shahmoradi and Ehtesham reviewed the improvements in prediction of survival of cancer patients by machine learning techniques in 2017. They placed more importance on analysis of gene expression data. Artificial Neural Network can be used to find the distinctive signature of gene expression in cancer patients. More clinical decision support systems can be developed by introducing the analysis of gene in the prevalent machine learning algorithms. They can reduce errors in estimation cancer survivability and provide treatment methods for individuals. Agarap compared the accuracy of six different machine learning algorithms on the Wisconsin Diagnostic Breast Cancer Dataset on 2018. The algorithms chosen were Linear Regression, GRU-SVM, Multilayer Perceptron, Nearest Neighbor search, Softmax Regression and Support Vector Machine. The comparison proved that all the algorithms were highly accurate for breast cancer detection. It can be concluded that the use of k-fold cross validation technique will provide more accuracy and also help in determining the most important parameters for the algorithms.

III. PROPOSED SYSTEM

In this paper we have used machine learning algorithms like Naïve Bayes, Decision Tree, Logistic Regression and Neural Networks algorithm to classify cancer patients and detect the type of cancer. Given a few parameters, our algorithms can predict whether the patient has malignant cancer or benign cancer.

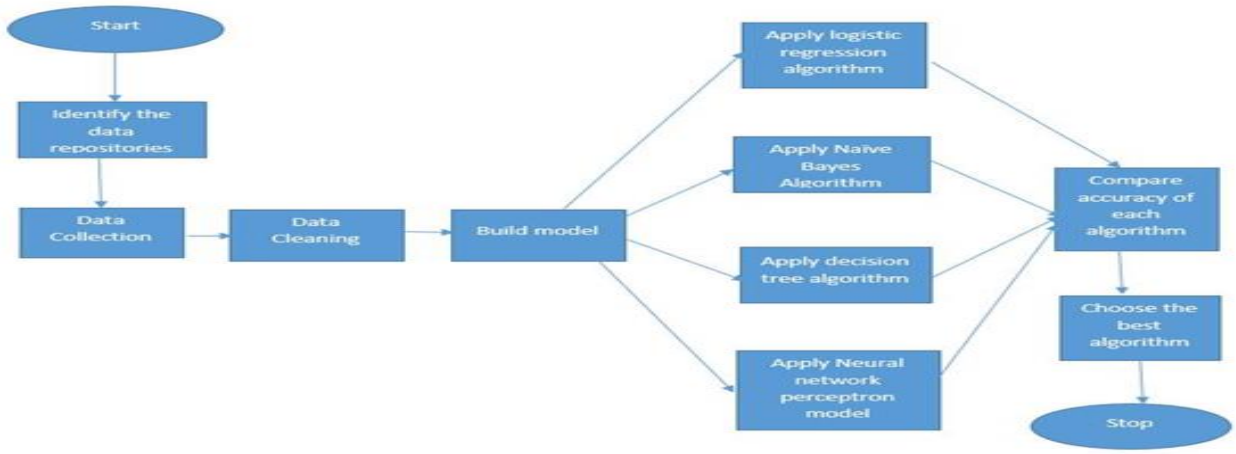

Fig 1: Proposed Methodology

IV. IMPLEMENTATIONS

A. Data File and Feature Selection

Breast Cancer Wisconsin (Diagnostic) Data Set from Kaggle repository and out of 31 parameters we have selected about 8-9 parameters. Our target parameter is breast cancer diagnosis – malignant or benign. We have used Wrapper Method for Feature Selection. The important features found by the study are: Concave points worst, Area worst, Area se, Texture worst, Texture mean, Smoothness worst, Smoothness mean, Radius mean, Symmetry mean.

B. Logistic Regression

The accuracy obtained is 97.48%.

```
> summary (model1)

Call:
glm(formula = diagnosis ~ concave_points_worst + area_worst +
    texture_worst + smoothness_worst, family = binomial, data = train_set)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.90252  -0.01041  -0.00048   0.00001   2.16216

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -47.355236  10.435561  -4.538 5.68e-06 ***
concave_points_worst  65.399367  21.795678   3.001  0.00269 **
area_worst          0.021814   0.004818   4.527 5.98e-06 ***
texture_worst       0.437942   0.112300   3.900 9.63e-05 ***
smoothness_worst    71.494642  28.911247   2.473  0.01340 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 538.504  on 409  degrees of freedom
Residual deviance:  41.826  on 405  degrees of freedom
AIC: 51.826

Number of Fisher Scoring iterations: 10

>
> t_pred = predict(model1,test_set, type="response")
> t=test_set['diagnosis']
>
> confMat <- table(test_set$diagnosis,t_pred>0.5)
> accuracy <-  sum(diag(confMat))/sum(confMat)
> confMat

    FALSE TRUE
  0    95    2
  1     2   60
> accuracy
[1] 0.9748428
```

Fig 2: Logistic Regression Output

C. Naïve-Bayes

The accuracy obtained is 98.4%

```
Naive Bayes

410 samples
  7 predictor
  2 classes: 'B', 'M'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 410, 410, 410, 410, 410, 410, ...
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE      0.9652921  0.9247277
   TRUE      0.9668365  0.9285583

Tuning parameter 'fL' was held constant at a value of 0
Tuning parameter 'adjust' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were fL = 0, usekernel = TRUE and adjust = 1.
> table(predict(model$finalModel,xtest)$class,ytest)
   ytest
     B  M
  B 98  2
  M  1 58
```

Fig 3: Naïve-Bayes Output

D. Decision Tree

The accuracy obtained using decision tree is 93.10%.

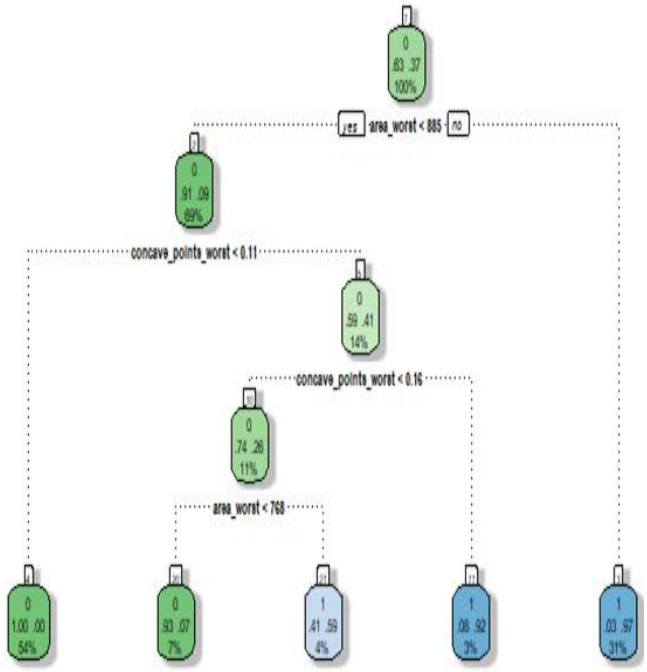

Fig 4: Decision Tree

E. Neural Networks

The accuracy obtained is 97.5%

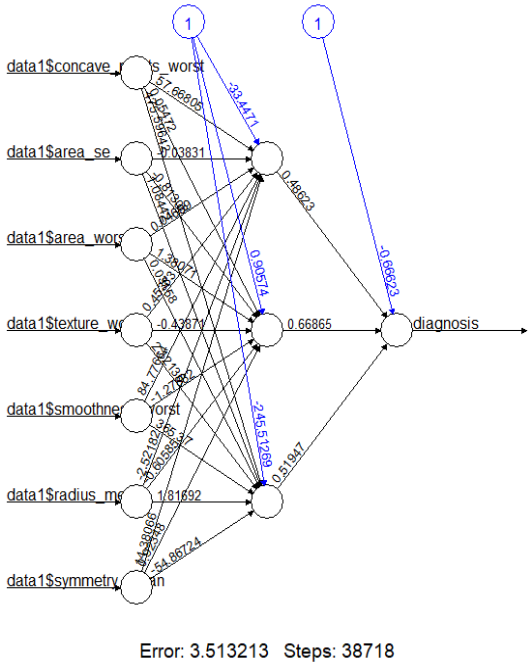

Fig 5: Neural Networks Model

V. RESULTS AND DISCUSSION

We have used Wrapper Method for Feature Selection. The important features found by the study are:

1. Concave points worst
2. Area worst
3. Area se

4. Texture worst
5. Texture mean
6. Smoothness worst
7. Smoothness mean
8. Radius mean
9. Symmetry mean

We have got the highest accuracy for Naïve Bayes algorithm- 98.4%. The accuracies obtained by both Logistic Regression and Neural Networks are close to 97.5%. The accuracy obtained using Decision Tree algorithm is 93.1%.

Results obtained for the Naive-Bayes Algorithm:

- Accuracy – 98.4%
- Precision- 98.9%
- Recall-98%
- F-score- 0.984

VI. CONCLUSION AND FUTURE ASPECTS

In this survey we have applied four different algorithms to cancer dataset and Naive Bayes was found out to be the most effective algorithm. Proper subset of features was found which was crucial in detecting malignancy. Integration of multi dimensional features can give more effective tools for detection of cancer. Other machine learning models like support vector machine, other models of neural networks (CNN or ANN) could be implemented. Other learning algorithms can be applied with using our chosen set of features. A dataset with more number of examples can be used.

REFERENCES

1. Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, Dimitrios I Fotiadis, Machine learning applications in cancer prognosis and prediction
2. Singhal, Ekta; Tiwari, Shamik. International Journal of Advanced Research in Computer Science; Udaipur Vol. 6, Iss. 1, (Jan 2015). Skin cancer detection using artificial neural network.
3. Rajendra Rana Bhat, Vivek Viswanath, Xiaolin Li, DeepCancer: Detecting Cancer through Gene Expressions via Deep Generative Learning
4. Azadeh BASHIRI, Marjan GHAZISAEEDI,* Reza SAFDARI, Leila SHAHMORADI, and Hamide EHTESHAM, Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review
5. Abien Fred Agarap, A Neural Network Architecture Combining Gated Recurrent Unit (GRU) and Support Vector Machine (SVM) for Intrusion Detection in Network Traffic Data.

AUTHORS PROFILE

Sreyam Dasgupta Completed B-Tech from VIT University, Vellore. He is going to study Masters in Data Science at University of Glasgow. He has published three papers – "Extended AES Algorithm with Custom Encryption for Government-level Classified Messages", "Image Compression using Bayesian Fourier" and "Smart Garbage Monitoring System" .

Ronit Chaudhuri Completed B-Tech from VIT University, Vellore. He has published research papers on "Image Compression using Bayesian Fourier" and "Smart Garbage Monitoring System". He is currently completed his 6 months' internship in Novartis, being awarded as the best techie.

Swarnalatha Purushotham is an Associate Professor, in the School of Computer Science and Engineering, VIT University, at Vellore, India. She pursued her Ph.D. degree in Image Processing and Intelligent Systems. She has published more than 60 papers in International Journals/International Conference Proceedings/National Conferences. She is having 15+ years of teaching experiences. She is a member of IACSIT, CSI, ACM, IACSIT, IEEE (WIE), and ACEEE. She is an Editorial board member/reviewer of International/ National Journals and Conferences. Her current research interest includes Image Processing, Remote Sensing, Artificial Intelligence and Software Engineering