# Machine Learning approach for Predicting Breast Cancer using Genomic Data

Saurabh Sharma
K.J. Somaiya Institute of Engineering & Information Technology, Sion, Mumbai
Department of Information Technology
University of Mumbai, India
sharma.sm@somaiya.edu

Neel Shah
K.J. Somaiya Institute of Engineering & Information Technology, Sion, Mumbai
Department of Information Technology
University of Mumbai, India
neel.shah3@somaiya.edu

Rishiraj Singh
K.J. Somaiya Institute of Engineering & Information Technology, Sion, Mumbai
Department of Information Technology
University of Mumbai, India
rishiraj.s@somaiya.edu

Reena Lokare
K.J. Somaiya Institute of Engineering & Information Technology, Sion, Mumbai
Department of Information Technology
University of Mumbai, India
reena.l@somaiya.edu

***Abstract*—Cancer prediction at an early stage is very crucial as the patient can then prepare for dealing with it. There are several Machine Learning models that help in predicting cancer by identifying samples of independent persons at high risk, facilitating the design and planning of cancer trials. These models use biomarkers like age, menopause, tumor-size, inv-nodes, breast, breast-quad dimensions to predict breast cancer. However, these models had major drawbacks of late prediction as well as low accuracy. So here presenting the system which uses gene expression profiles (genomic data) to predict breast cancer at an early stage. This model is built using different machine learning algorithms like a highly versatile support vector machine (SVM), Naive Bayes theorem, Decision tree and nearest neighbors approach to predict breast cancer using gene expression profiles.***

**Keywords—SVM (Support Vector Machine), Naive Bayes theorem, Decision tree, Nearest neighbors, Genomic data**

## I. INTRODUCTION

A major challenge facing healthcare organizations (hospitals, medical centers) is predicting the diseases with greater accuracy and at an early stage.
Here a system is proposed which will predict cancer of patients at its earlier stage by using genomic expression instead of only clinical expressions which will help us to achieve better accuracy. Gene data gives a better advantage since it has the potential ability to indicate cancer at an earlier stage which can be used to train the model more efficiently thus producing overall result more accurately. Different supervised learning algorithms like a highly versatile support vector machine (SVM) algorithm, Naive Bayes theorem, Decision tree and nearest neighbors approach to predict cancer of the patient are being used here. Using these methods, classification of patients will be done to predict whether a patient is suffering from cancer or not.

Over a long period of time, innovation on effective cancer treatment is in progress. Scientists applied different approaches such as screening at an early stage, in order to predict cancer type before the symptoms started to develop. The approach which was used by them was multi omics data viz biological data analysis. With the advancement of new technologies in the field of medicine, vast quantities of cancer data have been collected and are available for medical research. These datasets of new technologies are based on genomic data. However, the accurate prediction of a disease at an early stage is one of the most interesting and challenging tasks for physicians.

Gene Expression profiling is used in the proposed system. It is nothing but genomic data. It is the measurement of activity of 'n' number of genes at a single point of time to create a thorough picture of cellular function.

A Laboratory tool called a microarray helps in detecting various gene expressions simultaneously. The microscopic slides that have hundreds of tiny spots printed in specific positions are said to be DNA microarrays. Each spot in microscopic slides is known as DNA sequence or gene. The DNA molecules on such slides acts as probes that helps in detecting gene expression. These molecules are also known as transcriptome or RNA transcripts.

In this microarray analysis procedure, the RNA molecules of healthy individual and cancer patient are accumulated at one place. These samples are then converted into DNA samples of complementary version (cDNA). Each sample is labelled with different colors. The two accumulated samples are then combined on the microscopic slides. This process is called as Hybridization. After hybridization process scanning of microarray takes place by which expression of each sample or gene will be found. If the mutation of gene is greater than experimental sample, then the spot will turn red otherwise green. If the mutation is equal, then it turns yellow. In this way gene expression profile is generated.

A lot of research is being done on breast cancer. Researchers have developed breast cancer risk models which give probability of cancer occurrence. They make use of Clinical Data. There are few models which provide

such risk probability. International breast cancer intervention study model (IBIS), Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm model (BOADICEA), the BRCAPRO model and the Breast Cancer Risk Assessment Tool (BCRAT) also known as the Gail model.

IBIS and BOADICEA models are trained with around 19,000 samples and accuracy obtained by these models were 71% and 70% respectively. Whereas BRCAPRO and BCRAT models underestimated the risk and had an accuracy of about 68% and 60%.

Different methods are used to build such predictive models. Machine learning provides algorithms which can help in building such models. Machine Learning comprises different types of learning like supervised learning, unsupervised learning, semi-supervised learning etc. Supervised learning is used when the datasets consist of labelled output. Unsupervised learning is used when datasets does not have output label with it and semi-supervised learning is used when datasets consists of both labelled and unlabeled values. Here the datasets used to train models have labelled values, so a supervised learning method is used here. Different supervised learning methods are available and such methods are used for prediction. The prediction models built in this study is using Support Vector Machine (SVM), Naïve Bayes, Decision Tree and K-Nearest neighbors (KNN). All these are supervised learning algorithms.

## II. LITERATURE SURVEY

Jishnu Das et al. have compared various computational methods that have used different functional genomics datasets. They identify the molecular patterns that can be used for predicting prognosis of various human cancer tumors. Furthermore, they have outlined the challenges and how such approaches can be useful in solving those [1].

Cai Huang et al. have designed a software platform which predicts cancer from gene expression profiles. They used SVM based algorithm and for regularization they used Recursive Feature Elimination. Their main finding was that the model works best when it uses all probe-set expression profiles of individual patient tumors. They have achieved more than 75% accuracy [2].

Konstantina Kouroua Themis et al. have evaluated all the prominent available ML models. This includes ANNs, BNs, SVMs and DTs. This paper aims to validate the best approaches available so that they can be considered in everyday clinical practice [3].

Chaima Aouiche et al. have proposed a structure to identify stage specific cancer related genes by integrating multiple datasets. Also they have built a network by taking each sample pathway as vertices and relationships between genes as edges [4].

Qianfan Wu et al. have studied four articles that predicted cancer using genomic expression. These deep learning methods outperformed existing models such as prediction based on transcript-wise screening and prediction based on principal component analysis [5].

Esteva A et al. used Convolutional Neural Networks to classify skin cancer. They just used skin lesion images and disease labels to train the mode. The model showed great potential [6].

Russakovsky O et al. analyzed the past 5 years of Image classification competition and drew useful patterns and predicted the future development of image classification and its usefulness in disease prediction [7].

Hsu C-W et al. have explained in detail Support Vector Classification and its potential in disease prediction [8].

Gu Deshpande et al. all the currently used biomarkers for cancer prediction and concluded that these aren't enough. He then studied some more biomarkers which can increase the reliability of model if integrated with the existing biomarkers [9].

Sayes Y et al. have performed feature selection techniques by providing basic taxonomy of feature selection, discussing their use, and providing a variety of applications in both common as well as bioinformatics [10].

Radovic M et al. have proposed a temporal minimum redundancy-maximum relevance feature selection approach. The proposed system was able to handle multivariate temporal data without previous data flattening. Redundancy between the gene was computed using a dynamical time wrapping approach [11].

Gaul DA et al. have proposed a system using linear support vector machine. The results which were achieved provided evidence for the importance of lipid and fatty acid metabolism in OC and this can be used for clinical significant diagnostic tests. [12]

Guan W et al. have developed a system for ovarian cancer in which they developed new approaches for automatic classification of metabolic data. They have used SVM and cross fold validation technique which provided them highly accurate results [13].

Hoadley K. et al. have performed an integrative analysis using five genome wide platforms. In this paper methods such as Classification along with Correlation was used inorder to obtain better results [14].

Azuaje F. et al. have developed a model in which matching of tumor characteristics to the most effective therapy available and thus providing the patient with suitable precise medicine [15].

Salesse S. et al have Performed molecular mechanisms of leukemia induction to treatment of chronic myelogenous leukemia. In this paper they proposed a system with better accuracy [16].

Wenming Zhao. et al have provided a suite of genomic database resources. With the help of NGDC databases of genomic data a large number of requirements of data was made available publicly for study and research purposes [17].

## III. PROPOSED METHODOLOGY

The aim of proposed methodology is accurate prediction of cancer using genomic data. Cancer is a complex disease and complete causes behind cancer development are not yet fully discovered. Also, cancer treatment causes lots of expenses during treatment, and it increases as tumor grows. So by predicting cancer at an earlier stage, heavy expenses of medication can also be reduced.
The methodology of the proposed model is divided into four phases as shown in fig 1.

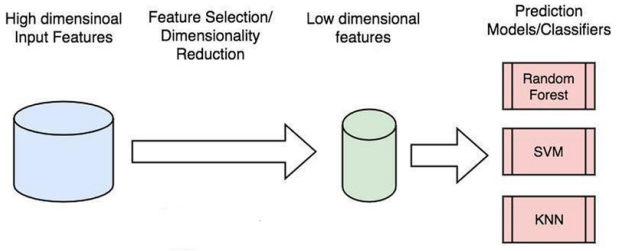


Fig 1 - Phases of prediction model

The phases are described below.

i) High Dimensional Input features -

Here microarray gene expression is extracted from online open source repositories [17-18]. The National Center for Biotechnology Information (NCBI) provides access to biomedical and genomic information. The datasets consist of 17,818 genes and 590 samples (including 61 normal tissue samples and 529 breast cancer tissue samples).

ii) Feature Selection/Dimensionality Reduction -

Since there are many genes, the model trained using all such genes may cause overfitting. Also, there are various genes which are not affecting the DNA mutation. To address this issue, major breast cancer causing genes are selected. There are 22 such major cancer causing genes namely BRCA1, BRCA2, ATM, BARD1, BRIP1, CDH1, CHEK2, MRE11A, MSH6, NBN, PALB2, PMS2, PTEN, RAD50, RAD51c, STK11, TP53, CASP8, CTLA4, CYP19A1, FGFR2, LSP1, MAP3K1 [19].

iii) Low Dimensional features -

The dataset having 22 dimensions is preprocessed first. All the field values are numeric values. However, there were many fields where the values were not present, so these values were replaced with mean values.

iv) Prediction Models and Classifiers -

After preprocessing, the dataset is obtained having 530 samples having 22 features (genes). Support Vector Machine algorithm is performed first on weka tool. This tool has various inbuilt machine learning algorithms. It also preprocesses the data and trains the model and plots various graphs. Initially, the dataset was passed to weka tool for model building. Later, SVM was implemented using python 3 on google colab to build the model. Along with SVM, Naive Bayes algorithm based model is also built using python 3.

## IV. RESULTS AND DISCUSSION

The results obtained from the machine learning algorithms are listed in table 1.

Table 1: Comparison of performance of Machine learning algorithms

| Sr No | Algorithm used | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|---|
| 1. | SVM | 0.9768 | 0.99 | 0.96 | 0.97 |
| 2. | Naive Bayes | 0.9259 | 0.94 | 0.91 | 0.92 |
| 3. | Decision Tree | 0.9898 | 0.96 | 0.95 | 0.96 |
| 4. | KNN | 0.9305 | 1.0 | 0.86 | 0.92 |

SVM is implemented on weka tool as shown in fig 2 and 63.42% accuracy is obtained. The model is tested as shown in fig 3. When SVM is implemented using python3, 97% accuracy is obtained. Naive Bayes algorithm was also implemented on python with an 92% accuracy along with KNN and Decision Tree algorithm 93% and 98% accuracy. In terms of predicting cancer, Decision tree and SVM both were the most accurate. Also for all four algorithms, their Precision, Recall and F1 score were calculated.

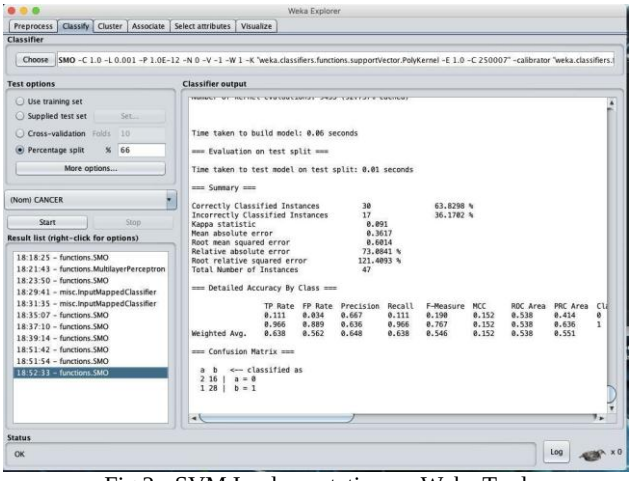


Fig 2 - SVM Implementation on Weka Tool

```
print("Expected:\n", list(test_set.CANCER[50:65]))
print("Observed:\n", list(test_set.Predictions[50:65]))
```

```
Expected:
 [0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0]
Observed:
 [0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0]
```

Fig 3 - Expected and Observed results for Cancer in SVM Model

From fig 3, '0' is label denoted for non-cancer sample and '1' is denoted for cancer sample.

Also genes namely BRCA1 located on chromosome 17q21 (5,6), and BRCA2, located on 13q12-13 (7–9), along with other 20 cancer causing genes inhibits growth of tumors. The microarray gene expression represents the mutation of genes, so if such genes are mutated then chances of tumor growing increases and eventually causing cancer. Thus due to such microarray gene expression early prediction of cancer is feasible.

## V. CONCLUSION AND FUTURE SCOPE

From the above study, it is clear that the cancer prognosis is possible in most cases using machine learning on high dimensional genomic data. Conventional cancer prediction models don't accurately predict cancer at an early stage. By using genomic data this void can be filled as it helps in early prediction. Genomic data integrated with advanced technology can be useful in other various medical prognosis as well. In this Application, four machine learning models for prediction of cancer are being implemented. However, this is a partial system. For early prediction of cancer, more dimensions of the individual sample may be required. These dimensions can be the lifestyle of the individual, hereditary etc. Acquisition of such dimensional datasets and combining it with gene expression can be the future task and based on such datasets machine learning models can be built.

## VI. REFERENCES


[1] Jishnu Das, Kaitlyn M Gayvert, and Haiyuan Yu "Predicting Cancer Prognosis Using Functional Genomics Data Sets" Published online 2014 Nov 2. doi: 10.4137/CIN.S14064 PMCID: PMC4218897 PMID: 25392695

[2] Cai Huang, Evan A. Clayton, Lilya V. Matyunina, L. DeEtte McDonald, Benedict B. Benigno,FredrikVannberg, and John F. McDonald, " Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy" Published online 2018 Nov 6. doi: 10.1038/s41598-018-34753-5

[3] Konstantina Kouroua Themis, P .Exarchosab Konstantinos, P. Exarchosa Michalis V .Karamouzisc Dimitrios, I .Fotiadisab " Machine learning applications in cancer prognosis and prediction " Published online doi.org/10.1016/j.csbj.2014.11.005 15 November 2014.

[4] Chaima Aouiche, Bolin Chen, and Xuequn Shang "Predicting stage-specific cancer related genes and their dynamic modules by integrating multiple datasets" BMC Bioinformatics. 2019; 20(Suppl 7): 194. Published online 2019 May 1. doi: 10.1186/s12859-019-2740-6 PMCID: PMC6509867 PMID: 31074385

[5] Qianfan Wu, Adel Boueiz,and Weiliang Qiu " Deep Learning Methods for Predicting Disease Status Using Genomic Data" Published online 2018 Dec 11 PMCID: PMC6530791 NIHMSID: NIHMS1024586 PMID: 31131151

[6] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542: 115–118. doi: 10.1038/nature21056 [PubMed]

[7] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vision. 2015;115: 211–252.

[8] Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification, Technical Report Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan, 2003

[9] Gu Deshpande and Ramji Rai An Overview of Prognostics Markers in Breast Cancer Med J Armed Forces India. 1999 Apr; 55(2): 129–132. Published online 2017 Jun 26. doi: 10.1016/S0377-1237(17)30268-X PMCID: PMC5531823 PMID: 28775603

[10] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23: 2507–2517. doi: 10.1093/bioinformatics/btm344 [PubMed]

[11] Radovic M, Ghalwash M, Filipovic N, Obradovic Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinformatics. 2017;18: 9 doi: 10.1186/s12859-016-1423-9 [PMC free article][PubMed]

[12] Gaul DA, Mezencev R, Long TQ, Jones CM, Benigno BB, Gray A, et al. Highly-accurate metabolomic detection of early-stage ovarian cancer. Sci Reports. 2015;5: 16351. [PMC free article] [PubMed]

[13] Guan W, Zhou M, Hampton CY, Benigno BB, Walker LD, Gray A, et al. Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. BMC Bioinformatics. 2009;10: 259–274. doi: 10.1186/1471-2105-10-259 [PMC free article] [PubMed]

[14] Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S. et al.Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158: 929–44. doi: 10.1016/j.cell.2014.06.049[PMC free article] [PubMed]

[15] Azuaje F. Computational models for predicting drug responses in cancer research. Brief Bioinform. 2016; pii: bbw065 (Epub ahead of print). [PMC free article][PubMed]

[16] Salesse S, Verfaillie CM. BCR/ABL: from molecular mechanisms of leukemia induction to treatment of chronic myelogenous leukemia. Oncogene. 2002;21: 8547–59. doi: 10.1038/sj.onc.1206082 [PubMed]

[17] Wenming Zhao, Yiming Bao, Shunmin He, Guoqing Zhang et al.(2020) "Database resource of the national genomics data center"

[18] Xie, Haozhe; Li, Jie; Jatkoe, Tim; Hatzis, Christos (2017), "Gene Expression Profiles of Breast Cancer", Mendeley Data, v1

[19] National Center for Biotechnology Information. Accessed on: Feb 13, 2020. Available: https://www.ncbi.nlm.nih.gov/guide/genes-expression

[20] Breastcancer.org. Accesses on: Feb 13, 2020. Available: https://www.breastcancer.org/risk/factors/genetics.