Cancer Prediction using RNA Sequencing and Machine Learning

ArunShalin L V[1] , Dr.I.Sharath Chandra[2] , Dr.Amol Dhakne[3] , Dr.R.Thiagarajan[4] ,Vaishali Rama Wadhe[5] ,Dr. T. Manikandan[6]

Assistant professor III,Department of Information Technology,Bannari Amman institute of technology, sathyamangalam , Erode[1]

Associate Professor and HOD-ECE,St.Peter's Engineering College, Hyderabad[2]

Associate Professor,Department of Computer Engineering,Dr. D. Y. Patil Institute of Engineering, Management and Research, Akurdi, Pune[3]

Associate Professor ,Department of IT,Prathyusha Engineering College,Chennai[4]

AssociateProfessor,K. J. Somaiya Institute of Engineering and Information Technology, Mumbai[5]

Professor,Department of Electronics and Communication Engineering,Rajalakshmi Engineering College,Chennai[6]

arunshalinlv@bitsathy.ac.in[1] ,sharath.inguva@gmail.com[2] , dhakne.amol5@gmail.com[3] , rthiyagarajantpt@gmail.com[4] ,vwadhe@somaiya.edu[5] , manikandan.t@rajalakshmi.edu.in[6]

Abstract:

Breast cancer is a type of tumour that develops in the breast tissues. It is the most common type of cancer found in women worldwide and is one of the leading causes of death in women. This article discusses Molecular subgroups are mostly used in research; they are not included in a patient's report as well as are not utilized to guide treatment. However, the application of subtype has vastly expanded; identifying tumour subtypes can extensions by determining which genes are expressed in tumour samples. Many investigators have worked on breast cancer diagnosis and prognosis; each technology has a distinct accuracy rate, which varies based on the situation, tools, and sets of data used. Our primary goal is to compare various existing machine Learning techniques using RNA sequencing in order to determine the best method for supporting large datasets with high prediction precision. The primary goal of this review is breast cancer detection, and this study gives all needed details to analyse machine learning techniques in order to gain a solid understanding of pattern recognition.

DOI Number: 10.14704/nq.2022.20.10.NQ55156 **NeuroQuantology 2022; 20(10):1729-1734**

Introduction:

There are numerous factors that raise one's likelihood of developing breast cancer; however, Genetic variants are frequently the underlying causes of breast cancer. Gene mutations that aid in cell growth and division, as well as mutations in genes that attenuate division of cells, tissue regeneration, and caspase activation, may also influence one's vulnerability to breast cancer development. Awareness of the disease and research grants have aided in the advancement of breast cancer diagnosis and treatment. Cancer survival rates have increased, remission rates have decreased, and the death rate associated with the illness has decreased. Recent time, the emergence of personalized medicine and genetic manipulation has the potential to change how doctors treat cancer. Personalized medicine is the generic term referring of medical care based on a cancer's cell membrane profile and the person's genome.

Breast cancer tumours are classified into four major subtype based on the expression levels by the cancer cells: Luminal B, Luminal A, HER2 Enriched, and Basal. Recognizing and researching these subtypes does have the potential to improve therapeutic intervention

*e*ISSN 1303-5150

and lead to the development of potential treatments. Currently, tumour stage, tumour grade, hormone receptor status, and HER2 status dominate prognosis and treatment decisions. Single - molecule subtypes are mostly used in research; they are not included in a patient's report and are not utilized to guide treatment. Nevertheless, vastly expanded; trying to identify tumour subtypes can improve prognosis by determining which genes are expressed in tumour samples.

The fingerprints of both the cancer and the nearby micro - environment are reflected in gene expression profiles of tumour tissues. As a result, single-cell transcriptome profiling distinguishes tumor-specific gene expression profiles from non-tumor compartments. We separated the breast cancer cells from the mixed population using cytogenetic patterns of gene expression. Data samples were used to match gene expression data across genomes.

Since the use of hormonal therapy for female hormones receptor (ER)-positive tumour forms, many genomic treatment options for breast cancer have indeed been investigated.
1. A successful gene-targeted medication[1] is genome-matched therapies of breast cancer to identify amplification of the human epidermal growth hormone receptor 2 genes HER2. Gene expression-based single - molecule subtyping is also widely used to aid medical decisions in tumours. Treatment for breast cancer alternatives have expanded thanks to targeted therapeutic approaches, which have improved significantly treatment outcome[2]. Nevertheless, in specific cancer sufferers, epigenetic and genotype expression profiling are typically used to characterise a bulk tumor, so even though cancers exhibit intratumoral heterogeneity that may affect the treatment effectiveness of a targeted therapies.

With high levels of genomic coverage, genetic variability in breast cancer has been revealed at a single-cell resolution[3]. Copy number abnormalities were discovered early in cancer formation and were stable, while single-nucleotide variants fluctuated greatly during tumour progression. A recent report6 on genetic changes in HER2-negative areas amongst HER2-amplified backgrounds revealed numerous genetic variants in a same tumour, implying a direct effect of genetic heterogeneity on therapy success[4]. Heterogeneity in gene expression also has a significant impact on patient outcomes. About 20% of ER+ cancers do not response to hormonal therapy or exhibit acquired methods to overcome later on. The existence of ER-negative cell in an ER+ tumour suggests a biological mechanism of treatment resistance.

Gene expression level variability in individual tumour cells may also result in rewired signalling automation allows through additional receptors for growth factors as well as a change in subtype from the source tumour in metastatic lesions[5].

Date Set:

Global gene expression attributes and attributes on 12553 gene mutations from 80 patients with breast cancer and three healthy adults are included. Medical evidence and numerous cancer categorisation from 105 women with breast cancer are included. This dataset was generated by using the subset decomposition method to remove noise from gene variables. The Comprehensive TCGA ID denotes a breast person with cancer; a few patients appear in both datasets. This is critical to our classification task because each patient has medical notes and transcriptomic data.

Proposed Method:

It is frequent in biological datasets for n, the quantity of observations, to be greater than p, the quantity of predictors. There is no more a unique weighted linear parameter estimation when p > n. When p>n, excessive variability and overfitting are frequently a serious problem. As a result, simple, highly regularised procedures frequently become the preferred ways. To resolve this challenge, numerous strategies such as Subset selection, Lasso and Ridge, and Dimensionality reduction may be used to remove unimportant parameters from a regression or dataset. We fit a factor that determines all 12553 variables that constrain or regularises the estimates, or equally, reduces the parameter estimates approaching zero, using the Lasso Shrinkage Technique. The lasso yields 58 non-zero coefficient genetic predictors. Those are the predictors used to categorise molecular tumour type; the code below identifies each predictor's position inside that dataset and obtains its column number.

Gene expressions were found with at minimum 99% of the data and saved in the data frame. As a result, the code below picks the predictors with non-zero coefficients from the data frame and saves them, together with the parameter for molecular tumour type PAM50 mRNA, in a different data frame.

Gradient boosting is a well-known machine learning approach for table data. It is strong enough to detect any nonlinear connection amongst your model goal and features, and it will handle missing data, anomalies, and large cardinality categorical values on your features. Gradient boosting is a variation of ensemble techniques in which numerous inferior models are created and combined to enhance overall performance. We would concentrate on the residuals from the initial step to enhance our forecast to achieve a better estimate.

We present an ensemble boosting approach for data analysis. We run the model and compute the accuracy; this is the best approximation due to the wide range.

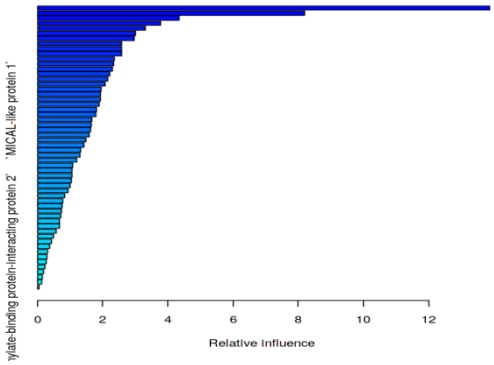


Fig.1. Relative Influence

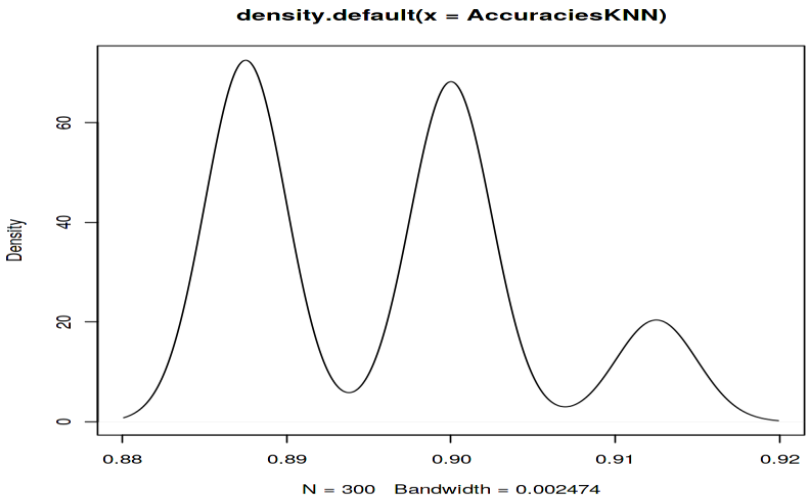
We propose an ensemble method with support vector machine and K – nearest neighbour classification

K-Nearest-Neighbors is a non-parametric categorization recognition algorithm. All variables must be normalised and scaled suitably before using KNN. To utilise KNN, first pick the amount of cluster neighbours (k), then the method computes a Euclidean distance using a weighted sum of the k nearest neighbours. Nearby values contribute greater to the aggregate than distanced values. KNN classifies data depending on how

similar it is to the training data set. One downside of KNN is that it strongly relies on human input; we must select the number of neighbours for categorization purposes. Because it is a non-parametric approach, group distance is immaterial because we are not focused on data distributions; moreover, we can utilise various data set.

Support Vector Machines are a generalised generalisation of a maximum margin classifier. They are designed for classification model where there are two different classes. This deliberate purpose, nevertheless, does not exclude employing the SVM approach for situations involving and over two classes. The optimal line separator is determined using SVM by locating the closest points in the convex hull, and a hyperplane bifurcates the nearest approach to an euclidean space. Based on borderlines vectors, the svm classifier identifies a test interpretation based about which side of a plane it is located. In order to perform better in identifying the residual data further away, the SVM approach allows certain observation to be on the erroneous side of the border and, in some situations, the inaccurate side of the hyperplane.

Gradient boosting is a method of machine learning for problems involving regression and classification with three major components.A error function that must be optimised.
Making predictions is difficult for a slow learner.To minimise the prediction error, an additive model is used to add weak classifiers.Gradient boosting generates a prediction model that is composed of a collection of weak prediction methods.

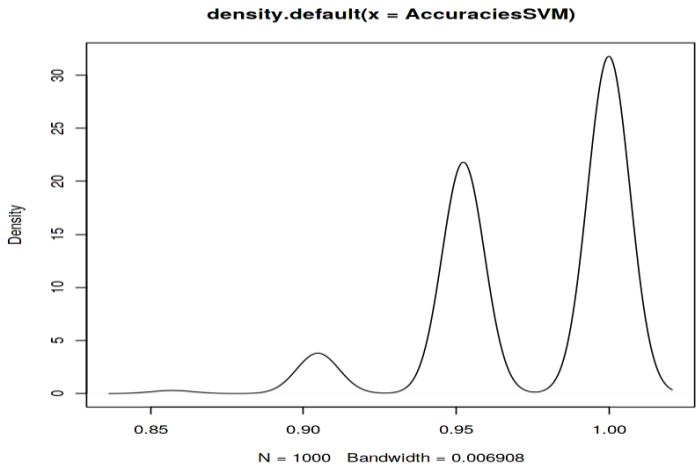


Accuracy with KNN

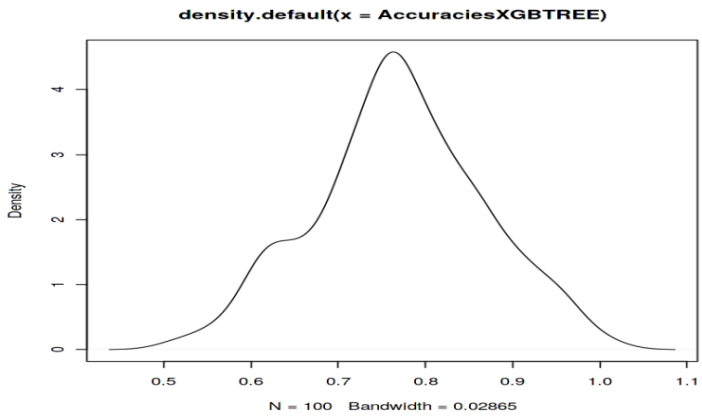
`hepatocyte nuclear factor 3-gamma` `hepatocyte nuclear factor 3-gamma` 13.87673019
`signal transducer and activator of transcription 6 isoform 1``signal transducer and activator of transcription 6 isoform 1` 8.20036773
`L-lactate dehydrogenase B chain` `L-lactate dehydrogenase B chain` 4.35111467
`39S ribosomal protein L40, mitochondrial` `39S ribosomal protein L40, mitochondrial`3.78091247
`protein LSM14 homolog B` `protein LSM14 homolog B` 3.30567083
`keratin, type I cytoskeletal 23``keratin, type I cytoskeletal 23`3.01010020
`dedicator of cytokinesis protein 1` `dedicator of cytokinesis protein 1` 2.95852944
`squalene synthase` `squalene synthase` 2.57653959

`receptor tyrosine-protein kinase erbB-2 isoform a precursor``receptor tyrosine-protein kinase erbB-2 isoform a precursor` 2.57352451

M-RNA sample



Accuracy with SVM



Accuracy with XGBoost

Classifier	Accuracy
K-Nearest Neighbour	90%
Support Vector Machine	98%
Ensemble XGBoost	78%

*e*ISSN 1303-5150

Conclusion:

Identifying and investigating the corresponding gene interaction pathways for each subtype of breast cancer might have a significant influence on tailored treatment for distinct patients.

We suggested adding the biological significance of regulatory information to differential expression analysis in this study. We used two machine learning methods and the XG boost ensemble methodology to improve the accuracy of the gene regulators. The performance has been shown together with the tumour and non-tumor gene prediction accuracy.

References:

[1] Higgins, M. J. &Baselga, J. Targeted therapies for breast cancer. J. Clin. Invest. 121, 3797–3803 (2011).

[2] Sparano, J. A. & Paik, S. Development of the 21-gene assay and its application in clinical practice and clinical trials. J. Clin. Oncol. 26, 721–728 (2008).

[3] Wang, Y. et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).

[4] Ng, C. K. et al. Intra-tumor genetic heterogeneity and alternative driver genetic alterations in breast cancers with heterogeneous HER2 gene amplification. *Genome Biol.* **16**, 107 (2015).

[5] Osborne, C. K. & Schiff, R. Mechanisms of endocrine resistance in breast cancer. *Annu. Rev. Med.* **62**, 233–247 (2011).

[6]R.Thiagarajan,N.R .Rajalakshmi ,M. Baskar ,P.Jayalakshmi "A Novel Solution for EconomizingWater by a Mix of Technologies with a Low Cost Approach",International Journal of Advanced Science and Technology Vol. 29, No. 7, April 2020

[7]Thiagarajan.R, Moorthi. M , Energy consumption and network connectivity based on Novel-LEACH-POS protocol networks,Computer Communications, Elsevier, (0140-3664), vol.149, pp. 90-98.

[8]R.Thiagarajan,Ganesan,Anbarasu,Baskar,Arthi,Rajkumar**,**Optimised with Secure Approach in Detecting and Isolation of Malicious Nodes in MANET**"** Wireless Personal Communication**,119**, pages21–35 (2021) **Springer** Jan 2021

[9] R.Thiagarajan, V.BalajiVijayan, Dr.S.Arun, I.Mohan**Novel Technique for Automation Billing in Smart Shopping**", International Journal of Scientific & Technology Research, Vol. 9 no.4, March 2020, PP:5363-5369

*e*ISSN 1303-5150