Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.47750/jptcp.2023.30.07.029

Machine Learning Algorithms for Breast Cancer Prediction

Senthil Kumar K M.E[1], A. Akalya[2*], J. L. Gayathri[3], V. Kanimozhi[4]

[1]Assistant Professor, Department of Information Technology K.S. Rangasamy college of technology Tiruchengode, India

[2,3,4]Department of Information Technology K.S. Rangasamy college of technology Tiruchengode, India

***Corresponding author:** A. Akalya, Department of Information Technology K.S. Rangasamy college of technology Tiruchengode, India, Email: akalyaamalraj6@gmail.com

Submitted: 16 February 2023; Accepted: 17 March 2023; Published: 08 April 2023

ABSTRACT

There are numerous subtypes of breast cancer, each with its own unique outlook. The evaluation of the expression of small gene sets is the primary focus of the current stratification methods. In the upcoming years, Next Generation Sequencing (NGS) is anticipated to generate a significant amount of genomic data. We investigate the application of deep learning, or machine learning, to the subtyping of breast cancer in this case study. We used pan-cancer and non-cancer data to create semi-supervised settings because there weren't any publicly accessible data. A wide range of supervised and semi-supervised designs are investigated with the help of Integrative omics data like microRNA expression and copy number variations.On our gene expression data challenge, accuracy results indicate that simpler models perform better than deep semi-supervised approaches. Deep model performance improves only marginally (if at all) when integrated combining several omics data types emphasises the need for additional research on bigger datasets of multi-omics data as they become accessible. In terms of biology, our linear model typically confirms. earlier classifications of gene subtypes. The development of a more varied and unexplored set of representative omics traits that may be helpful for subtyping breast cancer has resulted from deep methods, which imitate non-linear interactions.

Keywords: *Semi-supervised learning, deep learning, genomics, multi-omics, and variation auto-encoding*

INTRODUCTION

Because BRCA is known to cause a variety of lethal cancers, The use of molecular stratification has been successful. for predicting the clinical outcome of BRCA patients and assisting physicians in making treatment decisions over time. Unsupervised hierarchical clustering of BRCA microarray gene expression patterns was first used in the early 2000s. led to the discovery of the 'Her2-enriched', 'Basal,"

Normal-like, "Luminal A, “Luminal B, and 'BRCA intrinsic molecular subtypes. Since then, these subtypes have gradually gained clinically recognized prognostic value. Several Tests using multiple genes for prognosis have become an important part of BRCA patient management in recent years. Despite the consortium's Cancer Genome Atlas emphasis on the importance of various omics data in breast cancer taxonomies,

these tests only look at the expression levels of a few genes and do not take advantage of the emergence of large "omics" data generated by high-throughput technologies.

Deeper Learning

Artificial intelligence (AI) and machine learning have a branch called "deep learning" that aims to mimic human learning for specific sorts of data. In addition to predictive modeling and statistics, data science also includes deep learning. To further comprehend deep learning, think of a youngster whose name begins with "dog." By pointing to numerous objects and saying "Dog," the young child learns what a dog is and is not. The parent will then either respond, "Yes, that is a dog," or "No, that is not a dog." The youngster's awareness of the characteristics shared by all dogs grows as he continues to point to objects by constructing a hierarchy in which each level the abstraction to constructed the using of knowledge from the tier that comes before it, the toddler is, without realizing it, elucidating a complicated abstraction—the idea of a dog—through this process.

Genomics

It varies from "classical genetics" in that it takes into account all of an organism's genetic material at once as opposed to just one gene or one gene product. Genomic interactions such as epistasis, pleiotropy, and heterosis, as well as interactions between loci and alleles within the genome, are the main focus of genomics. Genomic research is now possible thanks to developments in next-generation sequencing methods and Fred Sanger's ground-breaking work. In the 1970s and 1980s, the lab of Fred Sanger set the standard for sequencing, genomic mapping, data storage, and bioinformatics research. This 1990s discovery led to the human genome project, a vast international effort that resulted in the 2003 publication of the full human genome sequence. The cost, volume, and speed of genome sequencing have all significantly increased thanks to next-generation sequencing technologies. Furthermore, many life-science databases and software programmers that support scientific study have been made possible thanks to advancements in bioinformatics. These databases collect and organize information so that it can be quickly accessed, compared, and analyzed. In the following sections of this course, we will look at various significant genomics resources.

Semisupervised Learning

A learning method called semi-supervised learning sits in between supervised learning (which exclusively employs labelled training data) and unsupervised learning (which uses no labelled training data). It is an example of poor supervision. Unlabeled data may increase learning accuracy when coupled with a modest amount of tagged data. significantly. To obtain labeled data for a machine learning task, it is usually necessary to employ a trained human agent, such as for transcribing an audio segment, or conduct a physical experiment, such as to determine the 3D structure of a protein or the presence of oil at a specific location.

Variational Auto Encoder

More specifically, our input data is Turned into an encoding vector, with each dimension representing a learned data attribute. The important thing to remember is that our encoder network produces a single value for each encoding dimension. The decoder network then attempts to replicate the original input using these values. A variation auto encoder (VAE) is a probabilistic representation of an observation in latent space. We will create a probability distribution for each latent property rather than a single value to represent each latent state characteristic.. A variation auto encoder (VAE) probabilistically characterizes an observation in latent space. Rather of producing a single value to represent each latent state feature, we will build a probability distribution for each latent property.

Literature Survey Support Vector Machines

Guyon, J. Weston, and colleagues proposed DNA micro-arrays, which allow researchers to examine thousands of genes at the same time and determine whether they are active, agitated, or

silent in normal or cancerous tissue. Because these new miniature exhibit systems generate a bewildering amount of raw data, a new sensible procedure should be developed to determine whether cancer tissues have the same gene signature as other types of cancer tissues or not. In this study, we examine the challenge of picking a limited group of genes from vast DNA microarray gene look data patterns. We develop a classifier that is suitable for hereditary diagnosis by employing readily available preparation examples from cancer patients and regular patients, as well as drug detection. Prior to attempting to address this issue, select genes using the association method. We propose a new method of gene selection based on the Vector Mechanism technique and Recursive Feature Elimination (RFE).

Breast cancer gene expression patterns identify TUMOR subclasses with clinical consequences. ROBERT TIBSHIRANI et al. propose in this paper. The tumors were classified into three types based on gene expression in differences: The ERBB2- overexpressing, basal epithelial, and similar to normal breasts. These groupings were quite robust, as demonstrated by clustering using two distinct gene sets: the first was a collection of 456 cDNA clones that were chosen initially to reflect intrinsic characteristics of the tumor; the second was a gene set that had a significant relationship with patient prognosis. According to survival analyses of a subset of patients treated uniformly in a prospective study and three fibro adenomas, the basal-like subtype of locally advanced breast cancer had a poor prognosis, and the estrogen receptor-positive groups had significantly different outcomes. This group includes 40 previously studied and described tumors. In total, 85 tissue tests from 84 people were examined. Tissue samples were stored in liquid N2 at - 170? C or -80?

This section looks at and provides information on the proper environment for Oncotype DX, MammaPrint, Prosigna, Endo Predict, Breast Cancer Index, Mammostrat, and IHC4 are a few examples of molecular tests. There is currently no multigene test that recommends adjuvant chemotherapy that is available for purchase. Importantly, the triple negative prognosis of the carcinomas is extremely variable, and emerging molecular markers that better comprehend this extremely heterogeneous subtype of the breast cancer can be improve by the outcomes of the therapeutic management by their illness. As long as traditional clinical, pathological, and immune histochemical indicators are used to guide treatment, the doctor may receive ambiguous results that call for additional testing. Those with luminal, HER2-negative, early-stage breast cancers who have up to three lymph node metastases are affected by this; the efficiency of adjuvant treatment in these patients is uncertain. [3].

Breast cancer risk prediction based on intrinsic subtype JOEL S et al. The intrinsic categories were significant for prognosis as independent Using intrinsic subtype and clinical data, a prediction of the model in the nodes having the negative breast cancer which has been developed. The combined model's C-index estimate (subtype and tumor size) outperformed either the clinic pathologic model or the subtype model alone by PCR has a 97% intrinsic subtype negative predictive value model indicated the success of neoadjuvant chemotherapy. Diagnosis by intrinsic subtype complements recognized for the breast cancer patients' prognostic and predictive markers. The continuous risk score's prognostic features will be useful in the treatment of node-negative breast cancers. Neoadjuvant treatment efficacy can also be predicted using the risk score and subtypes. [4] Prosing breast cancer gene signature test based on pam50 development and validation Prostigmata, BRETT WALLDEN et al. developed a On the Nano String nCounter Dx Analysis System, a subtype classifier and risk model based on PAM50 for breast cancer are designed for decentralised testing in clinical labs. 514 patient samples that were formalin-fixed, paraffin-embedded (FFPE) from each intrinsic subtype were used to train prototype centroids. Prototype centroids, generated during previous PAM50 algorithm training operations, were identified using hierarchical cluster analysis of gene expression data. A subtype-based risk model, known as Prosing ROR score, was developed using 304 formalin-fixed, paraffin-embedded (FFPE) patient samples from a clinical cohort with thorough annotation that did not receive adjuvant systemic therapy. Prior to

starting clinical validation investigations, the algorithm's prediction accuracy was tested using 232 samples from a group of patients who had been treated with tamoxifen. [5]

Existing System

In our circumstance, this is impractical for two reasons. For starters, it could have the unintended consequence of changing a cancer patient's current test data groupings, potentially negating the patient's human efforts in organizing her history.. Second, it has a significant computational cost because each new test data requires a large number of attribute test data group similarity computations. Existing ways to extract cancer illness prediction suffer from scalability, thus addressing this issue is critical. Cancer prediction connections are not uniform. This sort of information, however, is not always easily available in cancer prediction. Direct use of collective inference or label propagation would treat illness network links as if they were homogenous.

Proposed System

It has been demonstrated that the proposed disease prediction paradigm is effective in dealing with this prediction. The framework suggests a novel network classification strategy: To begin, use existing data mining techniques to classify based on the derived illness prediction, which is extracted from actors' latent affiliations via network connectedness. In the initial study, illness prediction was obtained using modularity maximization. Cancer prediction data has proven this framework's superiority to other representative relational learning methods. To extract sparse illness prediction, we suggest an effective edge-centric technique. We demonstrate that the suggested techniqueensures illness prediction sparsity. We investigate how search log signals like clicks and test data reformulations can be utilized to enhance search results. The dimensionality of the data affects how much data is needed to produce a reliable analysis in machine learning. The "curse of dimensionality" is a term used by Bellman to describe difficulties in dynamic optimization.

METHODOLOGY

It is established that the suggested framework for addressing this prediction, which is based on illness prediction, is effective. The framework suggests a novel strategy for classifying networks: first, use network connection-based illness prediction to discover actors' hidden connections; Second, classify networks based on the derived prediction by making use of existing data mining methods. Modularity maximization was used to extract illness prediction in the initial investigation. With the use of cancer prediction data, this framework's superiority over existing representative relational

learning methodologies has been demonstrated. In order to extract sparse illness prediction, we provide an efficient edge-centric method. We demonstrate that the sparsity of illness prediction under our suggested strategy is assured Test data reformulations and clicks and signals from search logs, such as test data reformulations, are examined in the context of assessing the applicability of test data groups. The quantity of data needed to make a reliable analysis in machine learning rises as the dimensionality of the data raises. When discussing issues with dynamic optimization, Bellman referred to this phenomenon as the "curse of dimensionality" A frequent solution to the high-dimensional datasets problem is to find a projection of the data onto a smaller set of variables (or features) that retains as much information as possible. A typical example of this kind of issue with small samples is data from microarrays. A high number of data points must be processed computationally because each data point (sample) can contain up to 450,000 variables (gene probes). Data from microarrays is a common example of this type of problem with small samples. A single data point (sample) can contain up to 450,000 variables (gene probes), and processing numerous data points is computationally intensive.

Pre-Processing And Data Visualization

The UCI Machine Learning Repository website allows users to download and save text files, including the Wisconsin Prognostic Breast Cancer dataset. The data is then saved as column headings with the associated properties, and a

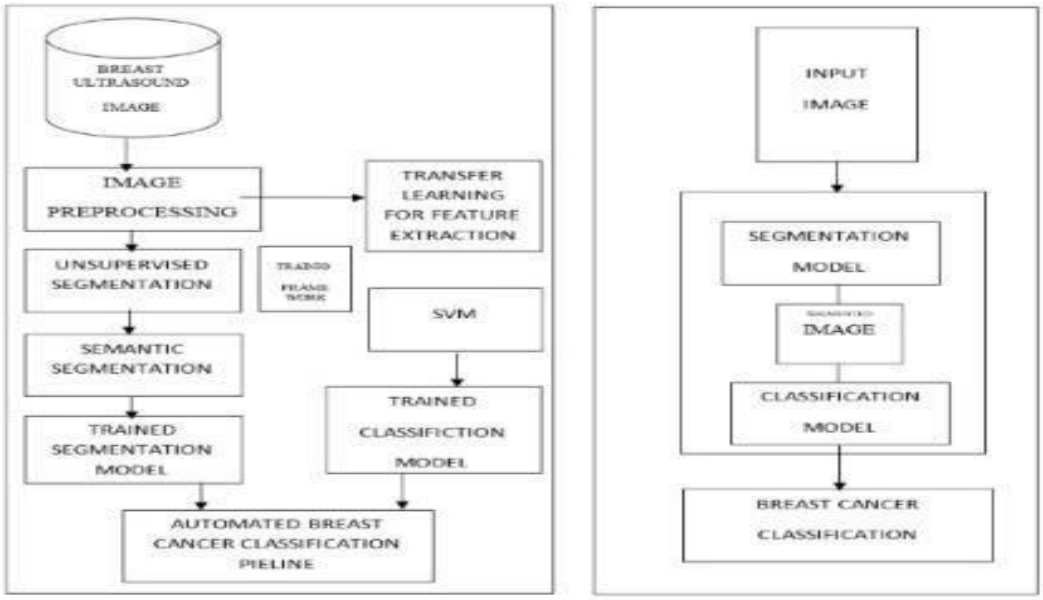
spreadsheet in Excel imports the file. The proper values are used to fill in the missing values. The performance of the classifier is unaffected by the patient cases' IDs. Thus, it is eliminated, and the result property describes the intended or dependent variable, bringing the number of characteristics in the feature set down to 33.

The next sections provide a detailed presentation of the computational methods used for feature relevance analysis and classification.

SVM-CFG Feature Selection Algorithms

The following is a general explanation of the supervised feature selection issue. We seek to identify a feature subset of size m consisting of the most informative features from a given data set (xi, yi), where xi Rd and yi 1, 2, c. Below is a quick description of the two feature selection techniques that performed well on the WPBC dataset.

Filtering by Mean and STD Score



For this module, the Wisconsin Prognostic Cancer dataset is taken from the UCI Machine Learning Repository and stored as a text file. The information is then imported into an Excel spreadsheet and saved as column headings with the corresponding characteristics. The missing parameters are replaced with appropriate values in this preprocessing state.

Fisher Filtering And Feature Selection Algorithms

The general supervised feature selection issue The following is an outline of SVM. The most informative feature subset with size m is what we are looking for. Below is a quick description of the two feature selection techniques that performed well on the WPBC dataset.

This module's filtering function ranks the input qualities in order of importance. A cutting rule allows you to pick a subset of these properties. In this cancer research field, it is necessary to identify both predictor and target features, distinguishing between recurrent and non-recurrent cases. The top-m features with high scores are selected based on their Fisher score calculation.

Svm Feature Reduction

A space with multiple dimensions is mapped into a space with fewer dimensions by using feature reduction. Feature extraction includes feature construction, sparse representation selection, and reduction of space dimensionality. All of these methods are often used as preprocessing for

prediction problems in statistics and machine learning, including pattern recognition. Even though scholars have been working on solutions to these issues for a while, feature extraction has recently attracted new attention. The feature space has significantly fewer features.

Scoring For Classification

Finding an useful subset of features when there are many descriptors for a certain issue area is a challenge for learning algorithms. Regression models with automated predictor variable selection are referred to as backward regression models. The logistic regression algorithm's iterations are broken down into the following steps.

The feature set with "ALL" predictors is the first step.

Step 2: Remove each predictor one at a time.

Step 3: "ALL" models with "ALL-1" descriptors are learned.

These iterations are repeated until the appropriate performance statistics (classification accuracy) are acquired or the predetermined goal size is reached. Following feature relevance, To categorise the various types of cancer cases in the Wisconsin Prognostic Cancer dataset, we employ twenty different classification algorithms.

CONCLUSION

In this study, we looked into a computational method for exploiting the larger datasets offered by high-throughput sequencing technology for subtyping breast cancer. challenge. To do this, we developed machine learning models that could effectively utilise the substantially bigger variable space, both supervised and semi-supervised. Identify unique breast cancer samples. Other data types and their combinations (miRNA and CNA data) showed that combining multi-omics Overall prediction accuracy did not increase as a result of the data, most likely due to the extremely high feature dimension in comparison to the number of samples offered. Nevertheless, single-layer Feed Forward Neural Networks (FFNN) and Variation Auto encoders (VAE) performed better than LR with combined multi-omics data.

REFERENCES

1. T. Srlie et al., "Gene expression patterns of breast carcinomas define tumor subtypes with clinical implications," Proc. Nat. Acad. Sci. United States, vol. 98, no. 19, pp. 10869- 10874, 2001.
2. F. Vieira and F. Vieira Schmitt, "An update on multigene predictive testing for breast cancer-emergent clinical signs," Front. Med., vol. 5, no. 248, 2018.
3. J. Parker et al., "Supervised risk predictor of breast cancer based on intrinsic subtypes," J. Clin. Oncology, vol. 27, pp. 1160-1167, 2009.
4. B. Wallden et al., "Development and validation of the PAM50-based prosigna breast cancer gene signature test," BMC Med. Vol. 1 Genomics Art. no. 54, 8, no. 1, 2015.
5. cancer Genome Atlas Network et al., Nature, vol. 490, no. 7418 (2012), pp. 61-70.