



AN IDENTIFICATION OF BREAST CANCER USING DATAMINING TECHNIQUES WITH ALGORITHMS

Shrikrishna S Balwante¹, Dr Mona Dwivedi²

¹ Research Scholar, Dept of Comp. Science and Engg, MGU, Bilkisganj, Sehore, MP.

² Professor in CSE, Dept of Comp. Science and Engg, MGU, Bilkisganj, Sehore, MP.

Email ID: balwantess@gmail.com.

Abstract

When it comes to significant neoplastic illnesses that affect women, breast cancer is by far the most common cause of death and is therefore often regarded as the deadliest of all malignancies. This is due to the absence of several early warning signs and symptoms, as well as the illness's aggressive aggressiveness. A molecular docking technique was also used to investigate the binding connection between proteins and curcumin. The Swiss model was used to make the proteins, and the PyMol programme was used to study them. ExPasy's ProtParam Proteomics service was used to explore the impact of the proteins' physicochemical properties. In addition, the SOPMA Self Optimized Prediction Method from Alignment server was used to investigate the proteins' secondary structures. Following that, each of them was subjected to an analysis using PROCHECK, ProQ, ERRAT, and Verify3D. During the first phase, the BSP-Slim server was used to dock the proteins made by breast cancer cell lines with curcumin. In this study, SVM, ANN, and Decision tree algorithms in breast cancer will be investigated, and this approach may be used to all types of malignancies.

7509

Key words: Data Mining, Classification Algorithms, Decision Tree, SOPMA, affected Images.

DOI Number: 10.14704/nq.2022.20.8.NQ44773

NeuroQuantology 2022; 20(8): 7509-7515

Introduction

Because of the rapid development of breast cancer, a huge number of women have perished, prompting scientists and researchers to perform a range of tests and studies on breast cancer diagnostics. Medical advances have proven that if the illness is detected and treated early on, the patient has a fair chance of surviving and recovering. Researchers may now access a large quantity of data that has been stored and made accessible because to advances in health and technology. A variety of procedures have been developed to aid in the detection of early signs and symptoms of illness. When a tumour is discovered, radiologists and other medical specialists must determine whether it is benign or malignant before proceeding with

therapy.

Machine learning enabled the development of a model capable of learning and identifying cancer based on past tumour diagnoses acquired from patients. This is now possible because to developments in machine learning. Over the recent decade, there has been an explosion of new automated varied classifiers, as well as enhanced methodologies and procedures for breast cancer analysis. You'll receive different results depending on your method of choosing. It is beneficial to test a wide number of models in order to identify one that is really accurate. Many different approaches and points of view have been explored on the categorization issue. Classification algorithms such as ANN and SVM are often employed to handle this



challenge.

The data utilized in this research has previously been used in other studies in the same field. In a research published in, SVMs that employ Radial Basis Functions (RBF) and Polynomial Functions (PF) as the classifier's kernel functions were proven to be 97.13 percent accurate. In the other two publications, artificial neural networks (ANNs) were employed (ANN). The initial publication employed an ANN back propagation strategy with genetic algorithms, resulting in a first-cleansing accuracy of 100 percent and a second-cleansing accuracy of 83.36 percent. When a feed-forward model and a back propagation technique with momentum and a variable learning rate are used, multi layer neural networks outperform single-layer neural networks.

Molecular docking is a valuable method for designing and developing rational medications, as well as researching their mechanisms of action. To do this, a molecule (ligand) is injected into the binding site of the DNA/preferential protein (receptor). Non-covalent techniques are used to produce a stable compound with greater efficacy and selectivity. The docking data may be utilized to calculate binding energy, free energy, and stability in complexes from several perspectives. The docking approach is currently used to provide preliminary predictions of ligand-receptor complex binding properties.

Basic Concepts

The pre-processing technique extracts the needed info from the big damaged picture using data mining tools. The pre-processing technique includes a number of phases such as data cleaning, data transformation, data integration, data reduction, and data separation. The act of reducing noise and ensuring that data is consistent and coherent is referred to as "data cleaning." The capacity to integrate previously unaccounted-for data and discover outliers is one benefit of this strategy.

Support Vector Machines (SVMs)

The support vector machine is a well-

known machine learning technique (SVM). It belongs to the larger class of machine learning methods known as kernel functions. Because of the usage of a kernel function in kernel techniques, non-linear conclusions may be obtained using linear methods. This enables them to categories data that does not seem to have a clear distribution. To divide the attribute space and generate a hyperplane, kernel techniques are required. It does this by classifying the data using support vectors, which are the instances that outline each hyperplane. Because of its ability to analyse high-dimensional data from a variety of sources, including gene expressions, SVM is frequently employed in biometric.

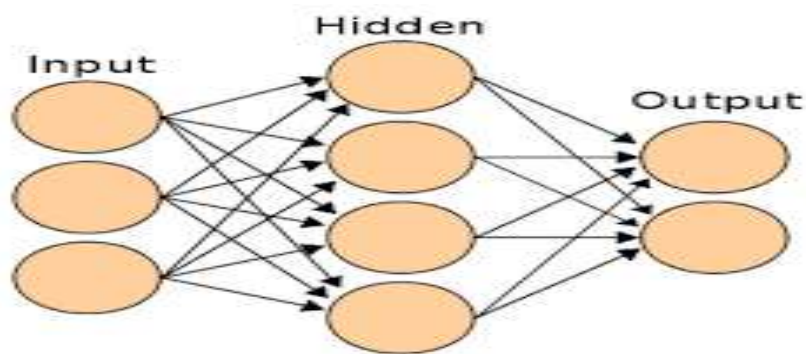
Based on their membership in that class, each individual in the data set is given a class. This is the algorithm's foundation. If there were many hyper planes, the one closest to the data point would be the best choice. As a result, the newly obtained data will be sent to the most appropriate grouping. This specific kind of classifier is known as a linear classifier.

Nerve Cells Produced Synthetically

The human brain network served as a guide for the creation of the Artificial Neural Network. ANNs may prove beneficial in the future. Most institutions had embraced artificial neural networks (ANNs) by the end of the 1980s for a variety of reasons. Neurons and weighted connections are the only physical components required for a neural network.

Input neurons, hidden neurons, and output nodes are the three kinds of neurons that may exist. Input neurons are neurons that take input from outside the network, while output neurons are neurons that generate output. The input and output neurons, on the other hand, are separated by a layer of hidden neurons in the network's center. Other neurons utilize the information created by the network's hidden neurons as input, which they receive as messages from the network's visible neurons. To generate an output, every neuron, excluding those that receive





and how they are linked to one another. Because neurons may construct numerous layers of this kind of network, multi layer perceptron (MLP) neural

information, analyses data from its neighbors. A variety of neural networks may be built depending on how neurons perceive input

nets are the most common. When MLP is utilized, neurons are organized in layers.

FIG 1 : ANN Flow Work

Neurons in the input layer, also known as the first layer, provide input to nodes. The last layer is the output layer, which is made up of neurons that convey messages to the outside environment. They represent the network's general inputs and outputs. There are one or more hidden layers between such two layers, each of which includes neurons that are concealed from view. Directed connections connect every node in the input layer to every node in the hidden layer.

Existing Methodology

The current method was tested with 5000 medically impacted photos and one non-medically damaged image. 2,500 photos of patients in varying stages of disease were supplied by medical institutions in AndhraPradesh. The participants in this study were chosen from the UCI Repository and the Heart Disease wards[5]. When employing NN without GA, heart disease identification accuracy increases by 12 percentage points,

Algorithm

and by 7 percentage points when utilizing GA + full training on affected pictures. KNN and genetic algorithms were unsuccessful for breast cancer or the first tumour.

A Proposed Strategy

The suggested approach employs support vector machines (SVM), artificial neural networks (ANN), and decision trees to improve the classification accuracy of breast cancer-affected pictures. Using the SVM algorithm on a big collection of photos from hospitalized patients. a pre-processing technique for detecting photographs with comparable issues Using SVM functions, we may extract a common attribute from a medically damaged image and then utilize ANN results from a decision tree technique to detect a cancer-affected image[4]. The suggested technique employs SVM, Ann, and a decision tree, using breast cancer as the detected variable.



Step 1: Load the medical affected image

Step 2: Apply pre-processing method on the affected image and Identify related affected image

Step 3: Related attribute to apply with SVM and preprocessing method from medical affected image

Step 4: common affected image from applying previous method

Step 5: applying svm and ann method

Step 6: Applying both SVM+ANN+ID3 with classified affected image and getting cancer affected image

Step 7: classified cancer affected image with diagnosis and prognosis of breast cancer affected image.

Description of Datasets

The UCI datasets was created through the collection of clinical samples. The information was organized in a database that mirrors the material's chronological categorization [7].

The group's samples were collected between January 1989 and November 1991. There have been 5000 occurrences of this in all. Each of the dataset's eleven attributes is explained in depth in the table below.

7512

Attribute	Type	Values
Sample Code Number	Numeric	No range
Clump Thickness	Numeric	1-10
Uniformity of Cell Size	Numeric	1-10
Uniformity of Cell Shape	Numeric	1-10
Marginal Adhesion	Numeric	1-10
Single Epithelial Cell	Numeric	1-10
Bare Nuclei	Numeric	1-10
Bland Chromatin	Numeric	1-10
Normal Nucleoli	Numeric	1-10
Mitoses	Numeric	1-10
Class	Nominal	2: benign 4: malignant



Statistical Analysis of the Data set

Tab. 2 shows the dataset's data study. The dataset's mean, median, SD, max, and min are presented[7].

Table 2: Analysis of the data set

Feature	Mean	Median	Standard deviation
Sample code number	1071704.099	1171710	617095.73
Clump thickness	3.418	3	3.816
Uniformity of cell size	2.134	1	2.051
Uniformity of cell shape	3.207	1	2.972
Marginal adhesion	1.807	1	1.855
Single epithelial cell size	4.216	2	4.214
Bare nuclei	3.845	1	3.844
Bland chromatin	2.338	3	2.338
Normal nucleoli	2.867	1	3.054
Mitoses	1.589	1	1.715

7513

Figure 2: SVM graph



Result Discussion

Using the correlation coefficient from table 3, we can choose the following attributes and compare their influence on the outcomes[7].



Table 3: Results of different features subset

Features Selected	SVM	ANN
All	97.1388 %	96.7096 %
7,4,3,8,2	96.5665 %	96.7096 %
7,4,3	95.422 %	95.7082 %
7,4	95.1359 %	94.7067 %
7	91.1302 %	89.2704 %

The confusion matrices for SVM and ANN allow for the observation that SVM provides a somewhat more accurate prediction of the Malignant class than ANN does.

Conclusion

In this research, two models were used to categories the UCI Diagnosis Breast Cancer data set. To construct the models, Artificial Neural Networks and Support Vector Machines are combined with feature selection and 10-fold cross validation. To conduct classification tasks, several combinations of sets of attributes were utilized, and the optimal parameters were discovered via data partitioning and analysis. The findings are compared to other categorization algorithms, such as SVM. SVM (97.1388 percent) outperformed ANN in sample classification (96.7096 percent). Furthermore, it was shown that both techniques had the highest accuracy when all accessible characteristics were used, suggesting that the features interacted and supported one another to help the whole system to a better conclusion. Future study should include additional testing utilizing alternative feature selection approaches, such as Info Gain, to increase classification accuracy.

Reference

1. Forouzanfar, M. H., Foreman, K. J.,

Delossantos, A. M., Lozano, R., Lopez, A. D.Murray, C. J., and Naghavi, M., 2011, "Breast and Cervical Cancer in 187 Countries between 1980 and 2010: A Systematic Analysis," The Lancet, 378(9801), 1461-1484.

2. Sarvestan Soltani A. , Safavi A. A., Parandeh M. N. and Salehi M., 2010, "Predicting Breast Cancer Survivability using data mining techniques," Software Technology and Engineering (ICSTE), 2nd International Conference, v2, 227-231

3. Hung, P.D., Hanh, T.D., Diep, V.T.: Breast cancer prediction using spark MLlib and ML packages. In: Proceedings of the 2018 5th International Conference on Bioinformatics Research and Applications, pp. 52–59. ACM (2018)4

4. Jaimini Majali, Rishikesh Niranjana, Vinamara Phatak, Omkar Tadakhe(2015)," data mining techniques for diagnosis and prognosis of cancer", International journal of advanced research in computer and communication engineering.pg.no.613-616

5. Miss.Jahanvi joshi ,Mr.RinalDoshi,Dr.Jigar Patel,"Diagnosis and prognosis breast cancer using classification rules", International journal of engineering research and general science volume 2,pg.no.315-323..



6. E.S.Samundeeswari (2015),”computational techniques in breast cancer diagnosis and prognosis: A Review” International journal of advanced Research ,pg.no:770- 775.
7. Reem Alyami,Reem Alyami, Investigating the effect of Correlation based Feature Selection on breast cancer diagnosis using Artificial Neural Network and Support Vector Machines

