Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 4): pp. 3975-3980
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.04.434

Comparative Analysis Of Breast Cancer Prediction Based On Machine Learning Techniques

Prakash Srivastava Department of Computer Science & Engineering, Graphic Era (Deemed to be University), Dehradun.

Dr. Suruchi Sharma Associate Professor, School of Management, Graphic Era Hill University, Dehradun.

Abstract

There are no typical symptoms of cancer; instead, it depends on the section of the body where the cancer has developed. Cancer is a well-known fatal disease that spreads throughout the body of the patient and is basically an uncommon, abnormal, or distinct sort of growth of tissues or cells. The goal of this endeavor is to make a contribution to the healthcare industry. It concerns the malignant or benign nature of a tumour. The report primarily focuses on the causes of breast cancer, using several machine learning algorithms that can be used for prediction, and then comparing the accuracy of each method. We utilised the most recent techniques because the earlier systems we evaluated applied fewer techniques, used an older database, and used overfitting of data so that we could compare more methods, we chose the most recent database. Our machine learning model has outperformed a number of current cutting-edge breast cancer prediction tools.

Keywords – SVM, Decision Trees , Random Forest, prediction.

1. Introduction

The fact that death rates have increased globally over the past few years—and that cancer remains the most common disease to cause those deaths—inspired us to take on this writing this research article. When a normal cell experiences fast mutations and transforms into tumors cells, cancer develops. There are several stages to it. When a mass is benign, it has not spread and is not cancerous. When a mass is malignant, it has spread to neighboring tissues and has become a true tumour. According to 2017 statistics, it is a sickness that does not even spare youngsters, as there were more than 15,000 [16] cases. Nearly 9.5 million people died from cancer in the world in 2018. Breast cancer is the most common type of cancer after lung cancer. Breast cancer affects primarily women. According to statistics, breast cancer claims the lives of almost 2 million women annually [17]. According to some estimates, breast cancer can strike a woman only once in her lifetime.

In addition to avoiding harmful substances like junk food, alcohol, smoke, and sedentary lifestyles, one should schedule routine check-ups. An early diagnosis can considerably lower the numbers. Men also experience this sort of cancer, in addition to women. However, given that this is a rare occurrence, the numbers are much too low when compared to women. A predictive model aids in producing fresh findings from previously understood raw data and offers a model that performs well with a certain type of processed data. In this study, we have divided the data into two previously identified classes using a variety of classification approaches.

Machine learning can therefore be quite useful in anticipating this kind of activity and accurately classifying it. The risk factors associated with breast cancer are shown in the figure below. This study attempts to apply three well-known machine learning methods (Decision Trees, Random Forest, and SVM) to the most recent database accessible through the UCI repository, evaluate the outcomes, and determine which method performs best with this type of data.

2. Related Work

Mehmet Fatih Akay [1] Support Vector Machine, which prioritised feature choice. They used the WDBC dataset from the University of California (UCI) machine learning repository for training the data and testing the trials. It was noted that the proposed technique launched and generated the highest classification accuracies ever attained in a model (99.51%, 99.02%, and 98.53%) for, respectively, 80-20% of (T-T partition), 70- 30% of (T-T partition), and 50-50% of (T-T partition).

Ronak Sumbaly et al. [2] proposed a method for diagnosing breast cancer using decision tree data mining. Their goal was to show how a lump appears and how decision trees could be used to determine and predict the cancer while using decision tree classifier using the Wisconsin Dataset and achieve 93 percent accuracy using the J48 dataset.

Shweta Kharya et al. [3] utilised this classifier to predict breast cancer and tried it on five distinct WBC datasets with various classifiers before using it on three different big datasets to determine the various accuracies it could produce. A different range of accuracy was attained, ranging from 80% to 93%.

With the WDBC dataset, Animesh Hazra et al. [4] used three techniques—Naive Bayes, ensemble algorithms, and SVM—to produce predictions based on five key classification criteria. As a result, they nearly attained a 95% accuracy rate across the board for all three approaches. The problem is in determining how well these strategies will work with large datasets.

Mandeep Rana et al. [5] employed two datasets, one for understanding the disease and the other for applying the recurrence, both downloaded from WDBC. They used SVM, KNN, Logistic Regression, and Nave Bayes as well as MATLAB. They were unable to put multi-class SVM into practice

3. DESIGN OF EXPERIMENT/ MATERIAL METHODS

The data was divided into two classes using a single train and test split using the built-in function train test split. The testing data made up 25% of the total while the training data was kept at 75% of the total, which means that when the missing value data from the dataset is removed, the training instances are kept at around 693, and 230 were kept for testing purposes. The accuracy of the decision trees technique is then somewhat increased by using a K-fold cross validation while keeping a value of K -10. There were a number of measures to be taken:

- Preprocessing: After importing the database, its properties are analysed, and any missing values are assessed and removed because they can cause inconsistencies in the findings later on. The data must then be normalised using a scalar and creating attribute values within a specific range.

- Learning: After the data has been divided into the two general categories of benign and malignant, various machine learning techniques must be used, and the system must then be ready.

- Post-processing: Following the analysis of the models' outputs and the evaluation of each one against a variety of performance indicators, we are better able to understand how various models perform against the provided dataset.

3.1 Decision Trees

Decision tree [9] is now the most widely used supervised machine learning algorithm, and it can perform both classification and regression. It is well recognized for its decision analysis, which employs a visual representation of a tree-like structure. It has a root node from which leaf nodes protrude. The central node of the tree displays the properties, while each leaf node represents a class.

$$E(T,X) = \sum_{c \in X} P(c)E(c) \qquad (1)$$

Where the tree's entropy is E(T,X), evaluated to determine homogeneity, its value of 1 indicates that the data is evenly distributed.

3.2 Random Forest

A more advanced variant of the Decision Tree method is Random Forest [10]. It is mostly used to solve categorization issues. As implied by the name, a forest is a grouping of many trees.

As a result, this technique is a collection of several decision trees that are applied to the dataset and make predictions. By casting votes for the top contender, the final decision is made. To prevent the sample from being overfitted, it employs the averages method.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (fi - yi)^2 \qquad (2)$$

Where N is data points.

3.3 Support Vector Machines

Advanced data clustering and classification techniques include Support Vector Machine [11]. It works with greater dimensional data (often at higher levels) and lots of features—really lots of features. These support vectors, known as hyperplanes, are found as tiger dimensional planes that divide the data into divisional planes. It divides the data into higher dimensions known as hyperplanes by using kernel tricks, which are frequently utilised in this. To categorize a point, cluster, or cluster of points that distinguishes diverse n planes from other hyperplanes and does so in an N-dimensional hyperplane.

Therefore, support vectors are those points that are in close proximity to hyperplanes. They are the ones that decide the length of the margin that should be between two hyperplanes and they also decide the position of the planes. Support vectors are also a deciding factor because they can change the position of planes if they are deleted.

4. Results and Discussions

The existing machine learning techniques viz., Decision Trees, Random Forest, and SVM were successfully operated on the dataset and comparative analysis is done on the basis of several parameters given below and it's performance is reflected in figure 1 and table 1.

Specificity: It is a measurement of the individuals who have been correctly diagnosed as not having the disease or, alternatively, the quantity of true negatives.

Specificity = Number of True negatives /Total number of healthy population

Sensitivity: It is a measure of the individuals who have been correctly diagnosed as not having the disease or, alternatively, the quantity of true negatives.

Sensitivity = Number of True positives / Total number of sick population

Accuracy: It measures how accurately or successfully the model(s) categorized the data.

Accuracy = Number of rightly classified outcomes/Total Outcomes

Table 1: Result Table

Classifier	**Accuracy**	**Sensitivity**	**Specificity**
Decision Trees	0.71	0.77	0.74
Random Forest	0.72	0.61	0.89
SVM	0.76	0.62	0.90

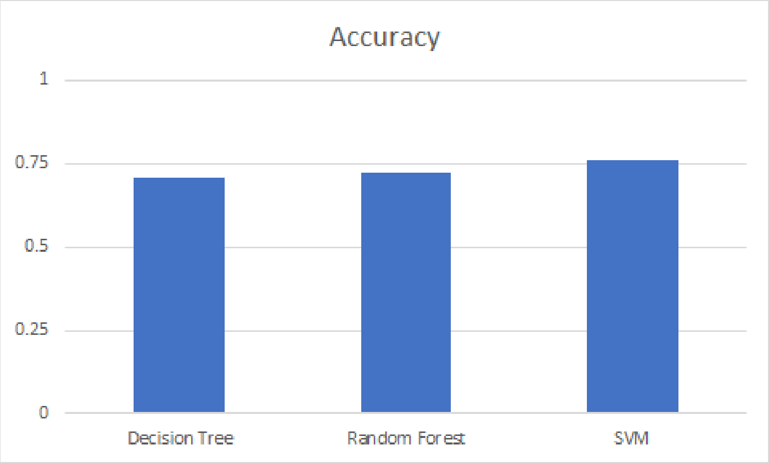
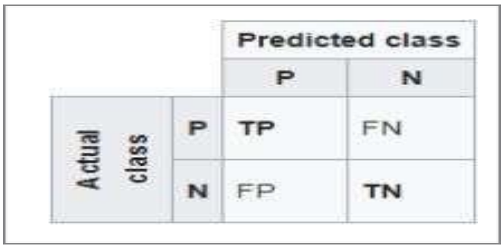


Figure 1: Comparative performance analysis of machine learning algorithms

4.1.Confusion Matrix

It is a matrix given in figure 2 which is used in machine learning where we can visualize different metrics and then evaluate our models based on those metrics' values. Here, we have two main classes: predicted class and actual class, which are two different classes of the predicted values and the real values, respectively. It can also be referred to as a performance matrix because it aids in determining how well models normally perform when using supervised learning algorithms.



		Predicted class	
		P	N
Actual class	P	TP	FN
	N	FP	TN

Figure 2: Confusion Matrix

5. Conclusions

We must advance our technology if we hope to treat breast cancer at an early stage. Consequently, using machine learning techniques enables us to do so. The goal of this work is to demonstrate how several machine learning methods can be utilized to simulate the real-time non-invasive diagnosis of breast cancer. Evidently, it produced seriously positive results in more than 8 out of every 10 cases that were evaluated. We may infer from all the findings that SVM has the highest accuracy, while decision trees have the lowest accuracy, or 71%, as shown in the results section.

References

1. Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications, 36(2), 3240-3247.

2. Sumbaly, R., Vishnusri, N., & Jeyalatha, S. (2014). Diagnosis of breast cancer using decision tree data mining technique. International Journal of Computer Applications, 98(10).
3. Kharya, S., & Soni, S. (2016). Weighted naive bayes classifier: a predictive model for breast cancer detection. International Journal of Computer Applications, 133(9), 32-37.
4. Hazra, A., Mandal, S. K., & Gupta, A. (2016). Study and analysis of breast cancer cell detection using Naïve Bayes, SVM and Ensemble Algorithms. International Journal of Computer Applications, 145(2), 39-45.
5. Rana, M., Chandorkar, P., Dsouza, A., & Kazi, N. (2015). Breast cancer diagnosis and recurrence prediction using machine learning techniques. International journal of research in Engineering and Technology, 4(4), 372-376.
6. Aruna, S., Rajagopalan, S. P., & Nandakishore, L. V. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer. Computer Science & Information Technology, 2(2011), 37-45.
7. Osareh, A., & Shadgar, B. (2011). A computer aided diagnosis system for breast cancer. International Journal of Computer Science Issues (IJCSI), 8(2), 233.
8. Bellaachia, A., & Guven, E. (2006). Predicting breast cancer survivability using data mining techniques. Age, 58(13), 10-110.
9. Quinlan, J. R. (1996). Learning decision tree classifiers. ACM Computing Surveys (CSUR), 28(1), 71-72.
10. Akar, Ö., & Güngör, O. (2012). Classification of multispectral images using Random Forest algorithm. Journal of Geodesy and Geoinformation, 1(2), 105-112.
11. Suthaharan, S. (2016). Support vector machine. In Machine learning models and algorithms for big data classification (pp. 207-235). Springer, Boston, MA.