May - June 2020
ISSN: 0193-4120 Page No. 4497-4502

Hybrid Predictive Model for Breast Cancer Detection

Bhavana.S[1], Bhavya.V.V[2], Charitha.S[3], C.Sonia[4], Aruna Kumara B[5]

[5]Assistant Professor, [1,2,3,4,5]School of C&IT
REVA University, Bengaluru, India
[1]bhavana0204@gmail.com, [2]bhavyavadde02@gmail.com, [3]charitha1598@gmail.com,
[4]soniarai.sona4@gmail.com, [5]arunakumara.b@reva.edu.in

Article Info
Volume 83
Page Number: 4497-4502
Publication Issue:
May - June 2020

Abstract
Cancer is a huge concern around the globe. This is a disorder that in many instances is deadly that has impacted many people's lives and will continue to impact many more people's lives. Breast Cancer is the second most cause of deaths in women. While cancer can be avoided and controlled in primary stages, an enormous percentage of patients are very late diagnosed. In one year, 40,000 women die from the disease, a woman died of the disease every 13 minutes. This is much harder to treat early breast cancer diagnosis. This paper presents a hybrid model which is a data mining technique to classify the smallest subset of characteristics that will guarantee a very reliable diagnosis of breast cancer as either benign or malignant in early detection. Naïve Bayes, Support Vector Machine and Random Forest classifiers are performed where they also calculate the time complexity of each of the classifiers. In this paper, the classification of Naïve Bayes is concluded as the best classifier with the lowest time complexity compared to the other two classifiers. Comparison of reliability of these three algorithms by precision, accuracy, recall and f-means, tests high comparison to the other classification algorithm. Such results are very favourable and can be used for diagnosis, prognosis, treatment and recuperation. The overall build hybrid model using ensemble method will be used to predict the cases based on the datasets with much higher accuracy.

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Keywords: *breast cancer, classification, complexity, navie Bayes, support vector machine, random forest, highest accuracy*

1. Introduction

The proliferation of breast cancer in breast cells is unregulated. Cancer develops as a result of mutations or irregular modifications in the genes responsible for controlling cell development and keeping them healthy. [1] This cell transition gains the power to keep dividing, creating more cells like it, and to form a tumour without any regulation or order. Calculating breast cancer level whether it is limited to one region in the breast or has spread across the breast or to other body parts to other tissues. [2] The word "breast cancer" refers to a malignant tumor formed in breast cells. Generally, breast cancer is either the beginning of the passages from the lobules to the nipples in the cells of the lobules which are the milk glands or the ducts. In the past, only three scientific elements, T, N, and M have been used to measure stage numbers. [3] the size of the cancer tumor and whether or not it has grown into nearby tissue (T),whether cancer is in the lymph nodes (N), whether the cancer has spread to other parts of the body beyond the breast (M) by TNM staging system describes the cancer cell size and how it had spread across. The patch of tissue, called the surgical margin or resection margin, has been checked before or after surgery to guarantee that it is free from cancer cells. [4] If there are cancer cells, decisions such as additional surgery and radiation will be impacted. After surgical

May - June 2020
ISSN: 0193-4120 Page No. 4497-4502

biopsy, lumpectomy and mastectomy, margins are monitored. The pathology report might state that there are surgical margins of Clear (also called Negative or Clean), Positive and Close type of cancer cells and the risk of surgery a woman has to undergo. Report on pathology can provide cell growth rate details [5] A higher percentage indicates a more aggressive, fast-growing cancer, rather than a slower, less aggressive cancer. The most common methods for diagnosing cancer among the current strategies are controlled machine learning.

Data mining is a collection of techniques and tools applied to the non-trivial method of collecting and providing tacit information from vast databases that was previously unknown, theoretically useful and humanly comprehensible. Mining of scientific data has excellent potential for discovering concealed trends in data sets. Data mining applications in medical and health science have proved successful and demonstrate significant potential for growth. Predictive and Descriptive of nature can be the model generated of data mining. A predictive model uses proven outcomes from specific data to make a prediction about data values. Predictive model classification methodology is utilized in this work. One of the most successful classification algorithms is Naive Bayes, SVM and Random Tree. And the rating of characteristics by assigning weights is the interesting concept to boost its efficiency. And Software is designed to accept the screening result of the woman and to forecast the risk of breast cancer in the future with greater accuracy. This paper demonstrates a method which is based upon hybrid mechanism. The hybrid method consists of three foremost classifier Naïve Bayes, SVM and Random whose accuracy of classifying is higher compared to other algorithms. Therefore, building a single hybrid model based on these algorithm leads to increases in overall accuracy of the prediction.

2. Related Work

Anand Sharma et.al, considered the studies on Breast Cancer Detection, it identifies benign from malignant breast lumps and Breast Cancer Prognosis assumes that Breast Cancer is expected to recur in women whose tumors have been excreted. Their researched-on improvement of data mining techniques concern to breast cancer diagnosis and prognosis and concluded that data mining strategies are a fantastic opportunity to discover trends concealed in the data that will help physicians making choices.

M Salehi et.al, inspected the adequate and active networks for breast cancer from clinically obtained data sets and Using different data mining methods, it is expected that the percentage of disease emergence will be estimated using the existing network. In this analysis, multiple architectures of neural networks are examined. The findings allow the patient to select a better diagnosis. The key component techniques are employed to solve a question of the large data set and the clustered complexity of the data, to decrease the data dimension and identify specific networks.[9]

Amit Kumar saxena et.al, proposed a hybrid classification model that involves a filter algorithm focused on correlation and a support vector machine as a classifier. This suggested classifier model refers to five high-dimensional binary type data sets. It is known that, in the case of three out of five high-dimensional datasets with a fairly limited number of characteristics, the suggested approach delivers better classification accuracy.[10]

Lily Wang et.al, proposed paper that describe the technical inconsistencies of previous work comparing random forests and the support vector machines and perform a new systematic evaluation of these limitations by the two algorithms. They used 22 diagnostic and prognostic data sets and proved that Random forests often overrun by a huge margin to support vector machines. Experiment results also underline the significance for the benchmarking and evaluation of bioinformatics algorithms of sound study research.[11]

R. Senkamalavalli et.al, considered the conventional medicine often remains issues such as poor accuracy and lack of self-adaptability. A classification algorithm Ada Boost SVM in conjunction with k-means is being suggested for early detection of breast cancer in order achieve an objective verdict, incomplete data, imbalance of data and other similar situations need to be taken into consideration, throughout this work. Calculating the accuracy, the confusion matrix, which offer physicians valuable signs for early detection of breast cancer, tests the reliability of the suggested approaches.[12]

Diana Dimitru proposed a paper on recurrent events on breast cancer using naïve Bayesian classification. The technique they have developed creative approach enables the automated processing of massive quantities of data related to breast cancer properties such that an optimal estimation of recurrent events is achieved. The precision of the study is about 74% which is compatible with the strongest performance obtained by other machine learning approaches. [13]

Woojaekim et.al, described about recurrence prediction model for novel breast cancer using Support vector machine in their paper. Retrospectively data were obtained from a Korean tertiary education clinic on 679 patients who undertook breast cancer operations between 1994 and 2002. The scientific, medical and epidemiological data forms were available. The precision was 99% provided the local tumour incursion attributes.[14]

M Sribala et.al, proposed a paper on efficient ensemble classifier for predicting the breast cancer at earlier stage. To examine the breast tissue and predicting that as cancer at earlier stage requires strong and efficient result for their diagnosis. Here they considered a classifier called ensemble to yield a better result. Engaging with proven classifiers including J48Naive Bayes, Random forest and SMO have applied ensemble

May - June 2020
ISSN: 0193-4120 Page No. 4497-4502

approaches for improving breast cancer estimation to identify breast tissues as carcinoma and fibroadenoma and concluded that the ensemble classifier Random forest bagging shown the highest accuracy of 83.65% and next better classifiers are SVM with stacking and Naïve Bayes with Ada boosting respectively. Ensured the higher accuracy of ensemble models than the provisional ones.[15]

Amit Gupta et.al, tried to search the trivial subset of features to guarantee that the breast cancer is marked as either benign or malignant with great accuracy. In addition to measuring the time complexity of each classifier, Naïve Bayes, Support Vector Machine and Ensemble classifier models are carried out and concluded that Naïve Bayes classifier provides average precision of 97.3978% and a time complexity of 0.102023 millisecond with just five dominant features ,compared to the other two classifiers this algorithm is less reliable. [16]

Yaochujin et.al, described about Black-box approaches that cannot clarify the treatment causes. A Random Forest (RF)-Based Rule Extraction (IRFRE) technique is built to circumvent this constraint by drawing up detailed and intelligible governing classification rules from a decision tree ensemble to diagnose breast cancer. First of all, the Random Forest number of decision tree models is designed to generate abundant decision rules. So, then an approach to rule-extraction is built to distinguish judgments from the qualified trees. Finally, an optimized multi-objective evolutionary algorithm (MOEA) is used to find an ideal rule predictor where the feature rule set represents the best compromise between accuracy and interpretability. [17]

MihailPopescu et.al, they used the National Inpatient Sample (NIS) results, accessible publicly through the Healthcare Cost and Utilization Project (HCUP), to train random forest classifiers to predict diseases. As the HCUP data is extremely uneven, we employed a dynamic approach focused on random sub- repeatedly. This methodology splits data from the testing into many sub-samples thus maintaining a total balance for each sample. Combining repetitive random sub-sampling with RF, we were able to solve the issue of class inequality and produce promising results. We estimated eight disease types with an average AUC of 88.79% using the national HCUP data package. [18]

3. Proposed System

The proposed method is designed using hybrid and ensemble method by using support vector machine, naive Bayes and random forest algorithm. We are using a hybrid model to increase overall accuracy of the prediction. These datasets are divided into test and training sets. The data from the training sets are trained using these algorithms. These three algorithms are the base learner for final model. These base learners are also called as weak classifier and used to generate the model for the strong classifier.

The processed data collected from base classifier is again trained on the test data sets such that the overall accuracy is increased. These base classifiers are then used to build the strong classifier. (Figure 1: Proposed Model)

A. SVM

The SVM stand for the support vector machine is a supervised learning algorithm. There are several theoretical reasons explain the superior empirical performance of SVMs in microarray data: e.g., they are robust to the high variable-to-sample ratio and large number of variables, they can learn efficiently complex classification functions, and they employ powerful regularization principles to avoid overfitting The algorithm is used for the binary classification and later on it was introduced with the help of kernel support. Basically, it used to create a hyperplane between the different dimensional data which is of completely two different classes. The underlying idea of SVM classifiers is to calculate a maximal margin hyperplane separating two classes of the data. The models that are having maximum space between the margin are considered as more efficient and preferred more as compared to the less margin space. SVM is highly used in the classification of the cancer cells. The algorithm will make a classification between the malign and the benign.

B. Naive Bayes

Naive Bayes algorithm is extensively used for the classification of large set of data. It has been using since long back for the classification of the document. We are using Naive Bayes with other two algorithm to reduce the error rate. Naive Bayes based upon the relationship of hypothesis to the prediction of the evidence. Suppose the class is given by C and the predictor is given by X then the probability of getting class to the given predicter is given by

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)}$$

P(C|X) is the posterior probability of class (C, target) given predictor (X, attributes). P(C) is the prior probability of class. P(X|C) is the likelihood which is the probability of predictor given class(X) is the prior probability of predictor.

C. Random Forest

Random Forests are being trained using bagging. Bagging or Bootstrap Aggregating is a random examination of preparing information subsets fitting a model to these littler informational indexes, and collecting expectations. Random Forests algorithm is trained using various methods one of them is bagging. Bagging or Bootstrap Aggregation consists of arbitrary testing of preparing information sub-sets, fitting a model to the little ones. This method makes repeated use of multiple instances for the training stage provided that we are sampling with substitution.

May - June 2020
ISSN: 0193-4120 Page No. 4497-4502

Tree bagging comprises testing subsets of the preparation set, fitting a Decision Tree when the bagging technique is applied thoroughly to the function space of the random forest it increases the randomness and the variety of the space. It randomly samples elements of the predictor space, instead of looking for the greedy approach which provides more diversity in the solution and also reduces the variance of the tree. Thus, this is termed as "feature bagging" and this is the feature that leads to the creation of a more robustmodel.

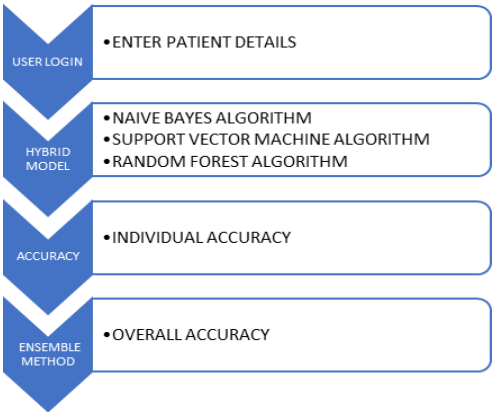

Figure 1: Proposed Model

The flowchart of the proposed hybrid and ensemble strategy is shown in Figure 2(Steps involved in prediction). Initially, the historical datasets are collected and then divided into training and test datasets. All the datasets are prepared for the pre-processing stage. Data pre-processingis a crucial step for any data analysis problem. It is often a very good idea to prepare your data in such a way to best expose the structure of the problem to the machine learning algorithms that you intend to use.

During the pre-processing stage data that is missing is replaced with the most current and optimized value. The features are assigned a certain range of value. These value changes dynamically as it is trained over the datasets. All the cells that are malign and benign are classified and then created a dataset on these categories. Then depending on accuracy of the algorithm a hybrid model is build and these prepressed data is passed down to these model.

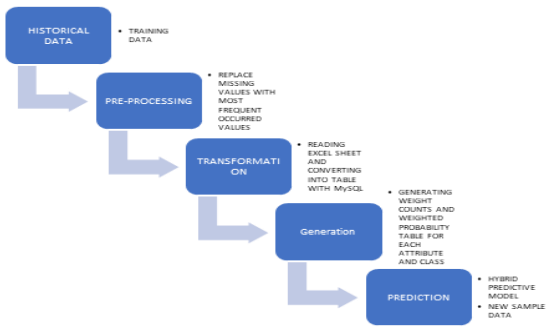

Figure 2: Steps involved in Prediction

A. Pseudo Code for Data Pre-Processing: -

%matplotlib inline import matplotlib.pyplot as plt #Load libraries for

data processing import pandas as pd #data processing, CSV file I/O

(e.g. pd.read_csv) import numpy as np from scipy.stats import norm

#visualization import seaborn as sns

plt.style.use('fivethirtyeight') sns.set_style("white")

plt.rcParams['figure.figsize'] = (8,4)

#plt.rcParams['axes.titlesize'] = 'large' data =

pd.read_csv('data/clean-data.csv', index_col=False)

data.drop('Unnamed: 0',axis=1, inplace=True)

#data.head()

B. Feature selection.

The cancer cell data contains various features which also makes one of the challenging task to develop a classification model. For clinical practice, the amount of usable cancer samples is very low relative to the number of characteristics, resulting in a higher risk of overfitting and classification performance degradation. Selecting features is a nice way to handle these challenges. The overfitting problem can be neglected by restricting the all considered feature space to a subset of features, hence reducing the difficulties resulting from a limited amount of sample and a large datadimensionality.

C. Classification methods.

As the sets of data are pre-processed, we evaluate the performance by predicting common classification methods formed by the unification of the three robust algorithm, against discrimination between normal samples and cancer cells. As we apply random forest, Naïve Bayes and SVM as the first stage and precision rate is noted down. And after this precision is used to build an overall high accurate classification model These algorithms have high accuracy in real-life applications. Then as for the final classification the ensemble method is used to generate overallaccuracy.

4. Result and Discussion

We have collected various data who are infected with breast cancer from the world. These datasets contain different attributes. Datasets are arranged according to their attributes. As Figure 3(Sample Dataset)represent some of the sample datasets we have used collected and used for training the models. The sample data consist attribute such as time which represent the year of the case. We have mentioned about the geographical location from which place the data is collected. Other attributes such as unit, value and type of the cancer tissue is mentioned. In our dataset ICD10 is also mentioned. ICD stands for International Classification of Diseases provide a wide variety of details about the disease such as signs,

May - June 2020
ISSN: 0193-4120 Page No. 4497-4502

symptoms, abnormal findings, complaints and social circumstances. The purpose of ICD is collection, classification and processing of different health related data. And ICD10 is revised and updated set of health-related data.

A hybrid model is build using the ensemble method for the prediction of the cancer cells. As the model is trained again and again it is improved very rapidly. The overall accuracy of the algorithm is increased by many times compared to the previously available models.

TIME	GEO	UNIT	SEX	AGE	ICD10	Value
2001	France	Number	Females	Total	Certain inf	5 157
2001	France	Number	Females	Total	Tuberculo	516
2001	France	Number	Females	Total	Meningoc	15
2001	France	Number	Females	Total	Viral hepa	439
2001	France	Number	Females	Total	Human im	248
2001	France	Number	Females	Total	Neoplasm	61 421
2001	France	Number	Females	Total	Malignant	58 262
2001	France	Number	Females	Total	Malignant	738
2001	France	Number	Females	Total	Malignant	699
2001	France	Number	Females	Total	Malignant	2 020
2001	France	Number	Females	Total	Malignant	5 786
2001	France	Number	Females	Total	Malignant	1 780
2001	France	Number	Females	Total	Malignant	1 705
2001	France	Number	Females	Total	Malignant	3 554
2001	France	Number	Females	Total	Malignant	4 624
2001	France	Number	Females	Total	Malignant	631
2001	France	Number	Females	Total	Malignant	11 088

Figure 3: Sample datasets

Table 1: Experimental Results

Algorithm's	Accuracy
NaïveBayes	85.11
SVM	96.28
Random forest	98.94
Hybrid Model	94

From the Table1 it is observed that, Naïve Bayes has accuracy rate of 85.11%, SVM has 96.28%, and Random Forest achieves 98.94% accuracy for the same dataset. Also the results shows that the proposed work achieves 94% accuracy. From the results it is observed that, Hybrid Model achieves better accuracy than the other accuracies.

5. Conclusion

It is know that breast cancer is one of the big problems. If the disease is not treated in early stages it can lead to the major issue. In some of the cases it even leads to the loss of the life. It would be easier to get treated if the disease is in the initial stages. If the disease is in early stages cost of treatment will also be less. Therefore, hybrid method is preferred more for detecting the disease at early stages. The cost of detecting the cancer cells is overall higher compared to other methods.

References

[1] Diana Dumitru "Prediction of recurrent events in breast cancer using the Naive Bayesian classification", 2009, Volume 36(2), Pages 92-96.

[2] Alexander Statnikov, Lily Wang and Constantin F Aliferis "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification" BMC Bioinformatics 2008

[3] R. Senkamalavalli and Dr. T. Bhuvaneswari "Improved Classification of Breast Cancer Data using Hybrid Techniques" International Journal of Advanced Engineering Research and Science (IJAERS), May 2018, Issue-5, Volume 5, Pg.no 77-81.

[4] Mohammed Khalilia, SounakChakraborty and MihailPopescu "Predicting disease risks from highly imbalanced data using random forest" BMC Medical Informatics and Decision Making 2011.

[5] Sutong Wanga, Yuyan Wanga, Dujuan Wang, Yunqiang Yin, Yanzhang Wanga and Yaochu Jin "An improved random forest-based rule extraction method for breast cancer diagnosis" Elsevier 2019

[6] Amit Gupta, AnimeshHazra and Subrata Kumar Mandal "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms" International Journal of Computer Applications July 2016, Volume 145, Pg.no 39-45.

[7] JareeThongkam, GuandongXu and Yanchun Zhang "AdaBoost Algorithm with Random Forests for Predicting Breast Cancer Survivability", IEEE , 2008.

[8] Shelly gupta, dharminderkumar and anand sharma "data mining classification techniques applied for breast cancer diagnosis and prognosis" Indian Journal of Computer Science and Engineering (IJCSE) ISSN: 0976-5166 Vol. 2 No. 2 Apr-May 2011 pg no. 188-195.

[9] M. Salehi, A. Soltani Sarvestani, A. A. Safavi and N.M. Parandeh "Predicting Breast Cancer Survivability Using Data Mining Techniques" 2nd International Conference on Software Technology and Engineering (ICSTE),2010, pg.no. V2: 227-231.

[10] Amit Kumar Saxena and Vimal Kumar Dubey "Hybrid Classification Model of Correlation-based Feature Selection and Support Vector Machine", IEEE, 2016.

[11] Subbalakshmi G. , Ramesh K., ChinnaRao M., Decision support in Heart Prediction System using Naïve Bayes, IJCSE ,2010.

[12] AlirezaOsareh, Bita Shadgar, Machine Learning Techniques to diagnose Breast Cancer. IEEE, 2009

May - June 2020
ISSN: 0193-4120 Page No. 4497-4502

[13] V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995.

[14] Woojae Kim, Ku Sang Kim, Jeong Eon Lee, Dong-Young Noh, Sung-Won Kim, Yong Sik Jung, Man Young Park, and Rae Woong Park "Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine" J Breast Cancer 2012 June; Volume 15(2):pg.no 230-238.

[15] M. Sri Bala, G. V. Rajya Lakshmi "Efficient Ensemble Classifiers for Prediction of Breast Cancer" International Journal of Advanced Research in Computer Science and Software Engineering, March 2016, Issue 3, Volume 6, Pg.no 5-9.

[16] Amit Gupta, Animesh Hazra and Subrata Kumar Mandal "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms" International Journal of Computer Applications • July 2016, Volume 145, Pg.no 39-45.

[17] Sutong Wanga, Yuyan Wanga, Dujuan Wang, Yunqiang Yin, Yanzhang Wanga and Yaochu Jin "An improved random forest-based rule extraction method for breast cancer diagnosis" Elsevier 2019.

[18] Mohammed Khalilia, Sounak Chakraborty and Mihail Popescu "Predicting disease risks from highly imbalanced data using random forest" BMC Medical Informatics and Decision Making 2011.